様式 C-19

科学研究費補助金研究成果報告書

平成２３年　６月　１日現在

機関番号：62601
研究種目：若手研究(B)
研究期間：　2008　～　2010
課題番号：20700228
研究課題名（和文）　マッシュアップを想定した複数教育コレクション提供サービスについての研究
研究課題名（英文）　Educational Materials Collection Services based on Mashups of Web APIs
研究代表者
江草 由佳（EGUSA YUKA）
国立教育政策研究所・教育研究情報センター・研究員
研究者番号：60413902

研究成果の概要（和文）：既存の Web API に基づくマッシュアップを想定し、論文検索システムを連携する仕組みについて検討した。特に、論文の同定について着目し、教育研究論文索引と CiNii との論文単位のリンケージについて検討し、調査を行った。調査結果で得られたデータを元に、プロトタイプシステムを試作した。また、マッシュアッ プ実験の 1 つとして、任意のテキストを対象に、そのテキストに類似した文書の検索を実現する手法「ふわっと関連検索」を開発した。

研究成果の概要（英文）： We studied a cooperative methodology of retrieval system for scholarly materials based on mashups through several existing Web APIs. We focused a method of article linkage between several resouces. Especially, article-level linkages of Education Paper Index (EPI) and CiNii were investigated. The results of linkages between EPI and CiNii were utilized for a prototype system of EPI as a hyperlink to an article indexed in CiNii. As another type of mashups, we developed a novel text retrieval method, Fuwatto Search, which is a document-by-document type retrieval method based on Web mashups.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2008 年度	1,500,000	0	1,500,000
2009 年度	1,000,000	0	1,000,000
2010 年度	800,000	0	800,000
年度			
年度			
総　計	3,300,000	0	3,300,000

研究分野：情報検索
科研費の分科・細目：情報学・図書館情報学・人文社会情報学　情報サービス
キーワード：マッシュアップ, Web2.0, 教育情報, サービス連携, 情報検索

１．研究開始当初の背景

マッシュアップとは、複数の異なる提供元の技術やコンテンツを複合させて新しいサービスを形作ることである。
たとえば、大手ネット書店サイト Amazon は Web API を公開し、書籍情報を利用者が自由に利用できるようにした結果、Amazon サイト上以外でも書籍情報を利用して、新しいインタフェースの提供、他社サービスとの結合による新しいサービス形態の誕生等、従来は行えなかったような検索サービスや、高付加価値サービスがうまれてきている。

一方で、教育コンテンツを含む領域では、商用サービスとは異なる領域での展開がおこなわれていることもあり、マッシュアップを取り入れたサービス方式の展開は見あたらない。つまり、現在の教育コレクションの提供状況は、提供側が用意した検索・提供画面からのみアクセスが可能で、いわば提供側のお仕着せの機能のみしか利用できず、利用者のニーズに応じてさまざまにカスタマイズして使用できるわけではない。
例えば、ある教育大学図書館が自館の蔵書検索で検索した結果と、教育論文データベースとを同時に検索できるシステムが欲しいと思ったとしても、容易に実現できる状況ではない。
マッシュアップを実現するためには、コンテンツ提供側は、他の機関が組み合わせて利用できるように（マッシュアップを想定して)、どのような値を渡せば、どのような結果を返すかを定義し、公開する必要がある。
しかし、教育コレクションについて、どのような値を渡せるようにするべきか、どのような結果を返せば、利用可能性が広がるかについての研究はされておらず、未知のことが多い。
また、受け渡す値の設定や、どのような結果を返せばよいかを検討するためには、それぞれのコレクションに対する情報ニーズの分析も必要となる。

２．研究の目的

そこで本研究では、複数種類の教育コレクションを Web-API を始めとする Web2.0 技術により提供し、利用者がより柔軟に再構成し利用できるようなサービス方法について、実際の複数種類の教育コレクション(教育研究情報センター・教育図書館の教育資料)を用いて研究する。

前提として、筆者は、教育資料の収集・提供を行なっている専門図書館に従事していることから、教育研究情報センター・教育図書館が管理・提供している情報つまり、研究成果報告書、教科書、教育研究論文索引、個人文庫等の特殊コレクションなどさまざまな資料を使った研究が可能である。

そこで本研究では、教育図書館が管理運営している様々な教育資料を元として、
提供主体から提供したサービスを利用者サイドや大学図書館のような利用者へ情報を伝達する情報サービス機関がさまざまな形で組み合わせて利用可能なマッシュアップを想定した提供サービスの方法について研究する。

３．研究の方法

既存の Web API に基づくマッシュアップを想定し、検索システムを連携する仕組みについて検討した。
本研究では、マッシュアップ方法として次の 3 つの異なる手法について検討した。

(1)既存の ID を用いる方法
(2)書誌記述から同定する方法
(3)テキスト類似検索を用いる方法

それぞれ、プロトタイプシステムの開発やデータ調査、システム評価などを行った。

４．研究成果

【既存の ID を用いる方法における研究成果】

既存の Web API にもとづくマッシュアップを想定し、図書館蔵書検索、論文索引検索とを対応付けるようなプロトタイプシステムを試作した（教育研究論文索引・検索（プロトタイプシステム）として公開している)。ベースシステムとして、2007 年度までの若手研究（スタートアップ)「図書館の情報提供システムにおける多言語アクセス：教育専門図書館を対象として」における研究成果として構築済の Z39.50/SRU/SRW による『教育研究論文索引』検索サービスを流用し、これの拡張またはこれと連携を図れるようなシステム構成でプロトタイプ試作した。具体的には Z39.50/SRU/SRW による『教育研究論文索引』検索サービスの検索結果から、他の情報サービス(国立教育政策研究所附属図書館の蔵書検索システム(NIER-OPAC)、Webcat、DOAJJ)へのリンクを付加して、当該の検索結果へ辿れる機能を追加した。
　プロタイプシステム構築にあたっては、既存のシステムと連携するにあたっての困難がないか、レコード形式に問題はないか、基本的な構築手法と構築時の課題を整理した。他のシステムと連携するためには、他のシステムのレコード同士をつなげるための仕組み(リンケージ)について検討する必要があることが分かった。もともと個別に作成・提供してきたデータベースには共通の ID が付与されていないことは多々あるため、同定するための仕組みについて検討することが課題である。プロトタイプシステムでは、ISSN を使って他のシステムとの連携をはかった。ISSN のない雑誌についての連携や、個々の論文ごとの連携について検討する必要があることが分かった。

【書誌記述から同定する方法における研究

成果】

教育研究論文索引(以下､EPI と呼ぶ)と CiNii との重複レコードを調査した。
ここで、重複レコードとは EPI と CiNii の両者に同一の論文が採録され、書誌レコードが取られていることを指す。
調査は 2 年度にわたって行った。教育研究論文索引データベースにおける調査対象はそれぞれ、2009 年 2 月 20 日までに登録されたデータ 154,624 件（第 1 期調査）と、2010 年 3 月 30 日までに登録されたデータ 164,643 件（第 2 期調査）である。
さらに、これらの全数調査に加え、第 2 期調査のデータを対象として、サンプリング調査もおこなった。

重複レコードの調査は以下の 3 ステップでおこなった。
1 レコード取得：EPI の全レコードを取得する。
2 候補検出：EPI と CiNii の論文書誌レコードを、書誌同定システムによって比較し、EPI 論文の書誌レコード 1 件に対し、CiNii の該当論文候補となる書誌
レコード 1 件を出力する。
3 判定：人手で EPI 書誌レコードと CiNii 候補論文書誌レコードを 1 件づつ比較し、同一論文に対する書誌レコードかどうかを判定する。

第 1 期、第 2 期の 2 つの調査を行い、EPI と CiNii と論文書誌レコードの調査を行った。人手判定を第 1 期は 99,500 レコード分、第 2 期は 17,000 レコード分を行い､EPI レコードのうち CiNii と重複して採録されているレコードは、164,643 件中、少なくとも 98,222 件(59.7%)あることが分かった。くわえて、1000 件をランダム抽出した調査では 63.4%の重複レコードがあった。

これらの調査結果は、EPI 上の論文書誌レコードから対応する CiNii 論文レコードへのリンキングサービスとして実サービスに反映し、利用者の利便性向上のための活用を図っている。
今後の課題として、CiNii 上の本文リンクの有無を考慮した検索表示や、さらに CiNii 以外の情報源へも同様のリンキングサービスを実装することを計画している。

【テキスト類似検索を用いる方法の研究成果】

「ふわっと関連検索」は、任意のテキストを対象に、類似文書検索を実現する手法である。以下では、この手法について説明する。

(1) 本文抽出
入力に PDF などのプレインテキスト以外の形式が指定された場合には、当該データの取得とテキストの抽出を行う。また、対象テキストとして Web ページが指定された場合は、当該ページを取得し、本文テキストを抽出する。この際、HTML タグを除く等の処理を行う。
(2) 特徴語抽出（テキスト中のキーワードの重み付け）
入力されたテキストまたは抽出した本文テキストをもとに、単語分割を行い、各単語それぞれのテキスト中での出現回数（tf）と、その単語の生起確率をかけあわせて、重み付けを行い、その重みを特徴語スコアとする。
(3) 検索クエリの発行
前項の処理で得られた特徴語群ベクトルから上位 n 件の単語を、論文データベースに検索クエリとして発行する。この際、検索結果がゼロヒットとなる特徴語は除外し、スコア順に下位の特徴語を順次、論文データベースに問い合わせ、データベース中から検索結果が得られる、n 件の特徴語リスト
を求める。次に、n 件の特徴語をすべて含む AND 条件式をクエリとして、論文データベースに検索をおこなう。この検索結果がゼロヒットとなる場合には、特徴語スコアの最下位 1 件を除外したうえで、n − 1 件の特徴語リストを AND 条件式として、クエリ発行する。以下、同様に、最終的に m 件以上の検索結果が得られるまで、漸次的に特徴語スコアの低い語から、AND 条件式に用いる特徴語を減らしていく。
(4) 検索結果の提示
前項で得られた、AND 条件式による検索結果を、詳細度の高い順に併合していき、最終的な検索結果ランキングとして提示する。
このような手順を採用した理由は、1) 最終的な検索結果がゼロヒットとなる確率を減らしてできるだけ多くの論文情報との出会いを生むことと、2) 詳細な文書類似検索が実装されていない論文データベースに
対しても単純なキーワード検索だけで簡易的な類似論文検索を実装することを意図したためである。

この手法は教育情報に限らず汎用的なしくみとしてマッシュアップ可能な手法であり、現在､教育研究論文索引､CiNii､CiNii 著者、NDL Porta、レファレンス協同データベース、J-STAGE、WorldCat、Springer、一橋大学 OPAC などに適用して、公開している。

また、この手法の評価のために、新聞記事に対して、評価実験を行った。評価実験の結果、新聞記事に対しては上位 10 件までの結果に

対して精度 0.25 を示し、平均精度（MAP）でも 0.17 の性能を示した。

今後は、これらの分析結果を踏まえた、特徴語抽出および類似度計算手法のさらなる改良や、他の文書種別に対する分析などを通じて、よりよい関連検索手法の完成を目指す。

図書情報、論文情報など Web API のマッシュアップ技術、適用例の情報収集のため、Code4Lib2010 カンファレンス、Code4Lib2011 カンファレンスに参加した。参加して得られた情報は、日本のコミュニティ での共有を図るために、報告会を開催した。報告会は Ustream で配信、録画も行い、Web において公開した。

さいごに：
オライリーが提唱した Web 2.0 の考え方を敷衍した``Library 2.0''とも呼ばれる
利用者参加型の図書館関連システムがいくつか登場しており、これらの図書館システムや蔵書管理システムにおいては、高付加価値サービスとなるサービスのバックエンドとして、OCLC Worldcat や各種 MARC データ、書店サイト Amazon など、大規模なデータプロバイダがサービス提供しているものも登場している。このように、技術的な課題については解決しつつある。

しかし、これらは商用サービスが先行したため、実際にどのようにサービスを提供していけばよいか、必要となるパラメータは何かといった核となるところはブラックボックスとなっている。

本研究は、実際に教育図書館で使用している教育コレクションという実在のデータと、現実の図書館サービスに関わりつつ、これらの核となる技術やサービスについて整理し、調査、システムを開発し、論文で発表することで、商用サービスで行なわれている技術を、人類の財産となる情報として遺せることが第一の結果と意義である。

現在先行しているサービスはどれも、全国書誌的で多くの一般的な図書情報を扱ったものであるが、本研究で扱うのは、教育という一主題に着目した点がユニークである。
全般的なサービスと主題を限定したサービスでは、必要となるサービスが変わってくる可能性があり、本研究を通じて教育分野という主題限定のサービスにおける知見が得られる。
また、教育という主題に限ってはいるが、教育分野には幅広い対象があるため、教育のなかでも様々な種類のコレクションがあり、さらに複数種類のコレクションに着目した研究としてもとりおこなうため、特定のコレクションのみに対して有効なサービスだけでなく、複数のコレクションに対しても有効なサービスとなるかどうかについても知見が得られる。
このように主題限定と複数コレクションをまとめてサービスする点について着目していることが本研究のユニークな点である。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔学会発表〕（計 3 件）
(1) “教育研究論文索引とCiNiiの重複率”; 江草由佳, 高久雅生; 情報知識学会 第 19 回（2011 年度）年次大会(情報知識学会誌 vol.21 no.2); 2011 年 5 月; 高松
(2) “簡易類似文書検索手法「ふわっと関連検索」の予備的評価と分析”; 高久雅生, 江草由佳;情報処理学会情報基礎とアクセス技術研究報告会; 2010-IFAT-99(14) 1-6;2010 年 7 月;東京
(3) “セレンディピティを促す論文検索ツール「ふわっと関連検索」”;高久雅生, 江草由佳;ディジタル図書館(38) 35-41;2010 年 3 月;東京
〔その他〕
ホームページ等

(1)教育研究論文索引・検索（プロトタイプシステム）
http://kaede.nier.go.jp/epi-search/sru-gw.rb
(2)ふわっと関連検索
http://fuwat.to/help.html
(3)Code4Lib2010 報告会
http://kaede.nier.go.jp/wiki/?code4lib2010
(4)Code4Lib2011 報告会
http://d.hatena.ne.jp/josei002-10/20110318/1300448576

６．研究組織
(1)研究代表者
江草 由佳（EGUSA YUKA）
国立教育政策研究所・教育研究情報センター・研究員
研究者番号：60413902
(2)連携研究者
高久雅生（TAKAKU MASAO）
物質材料研究機構・科学情報室・主任エンジニア
研究者番号：00399271