D-028

異種分散情報源の統合による書籍情報の統括的検索の実現

Implementation of Unified Search of Book Information by Integrating Distributed Resources

金子　護†　　成　凱‡
Mamoru Kaneko　Kai Cheng

1．はじめに

インターネットの普及に伴い，図書館蔵書情報、図書販売情報を含むさまざまな情報源がインターネット上で公開され、必要な書籍情報が簡単に入手できるようになっている．一方、これらの情報源が他の情報源との整合性はあまり意識されず、個別に管理運営されているので、複数の情報源を統合的に利用できない問題点がある．例えば、大学で図書購入を検討する際には、まず図書館の蔵書状況を検索し、当該書籍が図書館に所蔵しているかを確認する．所蔵していない場合は、次に図書販売サイトを検索し在庫状況や価格等を確認して購入可否の意思決定を行う．大量の図書を検討する際に図書毎に同じ操作を繰り返さなければならないので煩わしく、また効率も悪い．このため、近年異種の情報源を適切でかつ効率よく統合し統合的な検索サービスを提供するニーズが非常に高まっている[1].

従来、データベース分野ではスキーマ変換などの手法を用いて情報源そのものの統合を目指していた．しかしインターネット上で個別の組織や個人によって管理される分散情報源に対しては、このアプローチは有効ではない．異種分散情報源を統合するために、情報源そのものの統合ではなく、情報源への検索結果の統合が必要不可欠である．

本研究では、我々はネットワーク上で分散されている異種の情報源の統合により統括的な書籍情報検索を実現するとともに、異種分散情報源統合の有効性を検証する．具体的には、図書販売サイト(本研究では Amazon を利用する)、図書館蔵書検索サイト(OPAC) をマッシュアップさせ、複数の図書情報サイトの情報を統括して検索できるシステムを開発する．

2．異種分散情報源の統合

2.1　異種情報源統合の枠組み

異種情報源統合の一般的枠組みとしては、**基盤層**、**仲介層**、**アプリケーション層**に階層化することができる[1]．基盤層はデータベースをはじめとする様々な情報源そのものの統合を目指す．アプリケーション層では異種情報源の統合による利用者の問題解決や意思決定の支援を目的とする．例えば、利用可能な図書館蔵書情報と図書販売情報を統括的に検索することで図書購入の意思決定を支援する．仲介層は、基盤層とアプリケーション層の中間に位置付き、利用者と情報源との仲介を行う．インターネット上で分散された異種情報源に対しては、基盤層での統合より、仲介層、アプリケーション層での統合が有効である．

仲介層は一般的に以下のモジュールからなる：(1)特定の問題領域における異種情報源の統合を目的とするメディエータ(2)利用者と情報源間の効率的でかつ柔軟な情報配送を重視するファシリテータ．メディエータは情報源の変化に対する頑健性を保ちながら、複数の異種情報源へのデータアクセス、データ同一性の判定、類似データの統合などを行う．異種性を対処するために情報源ごとにラッパーを用いることが一般的である．また、ファシリテータは必要に応じて情報源やメディエータが提供できる情報に関するメタ情報を管理し、利用者の要求とのマッチメイキングと効率的な情報配送を支援する．図 1 では異種分散情報源統合における仲介層の枠組みを示している．

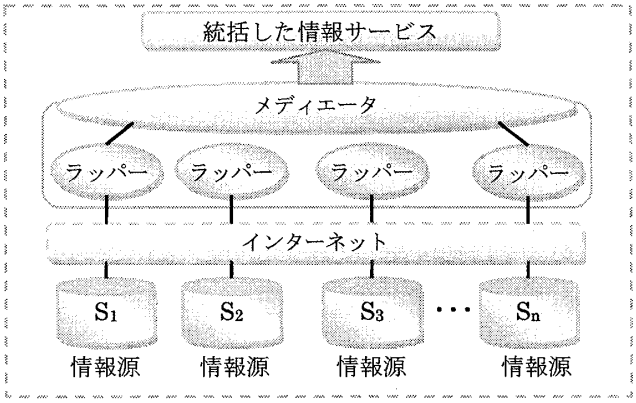

図 1 異種情報源統合における仲介層の枠組み

2.2　書籍関連の異種情報源

図書購入の意思決定を行う際に、図書館の書籍情報と複数の書籍販売サイトからの商品情報を統合した検索が望まれている．しかし、情報源によって提供される情報の種類やインターフェースが異なり、統合的検索の実現は簡単ではない．特に情報源のインターフェースは以下のいずれかが提供される．

1. UI：人間利用者向けのインターフェース
2. API：アプリケーション向けのインターフェース

大手ネット書店 Amazon 社では通常の書籍検索の UI とウェブサービスとしての API 両方を提供している．API を通して Amazon ウェブサービスを通して、アプリケーションは Amazon 書籍検索を利用でき、得られた情報も確実で精度が高い．一方、他の多くの図書販売業者や図書館においてはアプリケーション向けの API がなく、人間利用者向けの UI しか提供されていないため、アプリケーション層ではこれらの情報源は直接利用できず、得られる情報が確実ではなくて精度も低い．

インターフェースの異種性を統一するために、**ラッパー**を用いるのが一般的である．ラッパーとは、あるプログラムのインターフェースを変換して別のプログラムで利用できるようなインターフェース提供するプログラムのことである．例えば、ラッパーは人間向けのインターフェースしか用意されていないプログラムを包み、アプリケーションのなかで利用できるような API を提供する．

† 九州産業大学大学院　情報科学研究科
‡ 九州産業大学　情報科学部

インターフェースの異種性に加え、複数の情報源より得られる情報の種類、形式、精度等も異なるため、統合する前に、一連の処理が必要不可欠である．例えば、書名、著者目、出版社等の情報が一部欠けていたり、他の情報と混在していたりすることがあるので、同一書籍であるかの判断は難しい．

2.3 書籍情報の同一性

複数の情報源から得られた書籍は同一のものであるかどうかを判断することが図書情報の統括に必要不可欠である．書籍情報に、書名、著者、出版社 、出版国 、価格、出版日、ISBN、ページ数等がある．その中に、ISBN のような的確で一意性のある項目もあるし、書名、著者等一意性のない情報もある．一意性のある項目とは、唯一の図書を特定でき、同じ値は唯一の書籍しか持たない項目を指す．ISBN のような一意性のある項目がわかれば、書籍情報の同一性を判断するのが容易であるが、全情報源においてそのような項目が提供されるとは限らない．書籍情報の同一性を判断するために、以下の方法を用いる．

(1)近似的な同一性判定

一意性のある項目を持たない情報源において、情報の同一性を判定するために、複数の項目を用いて類似度を計算する．例えば、題名、著者、出版社、出版日のそれぞれの同一性を計算し、これらの結果の合計で書籍情報の類似度とする．類似度はある閾値を超えた場合、同一書籍と判定する．この方式は共通の項目があると適用できるので適用範囲が広い．しかし、正確な判定ではないので、誤判定(false positive と false negative)が生じる可能性がある．

(2)正確な同一性判定

一意性のある項目が提供された場合は、正確な同一性判定ができる．例えば、ISBN によって同一書籍を特定する．この方式は確実に同一性を判定することが可能であるが、全ての情報源に提供されているとは限らない問題点がある．

3. 統合的書籍情報検索の実現

3.1 書籍情報の統括的検索

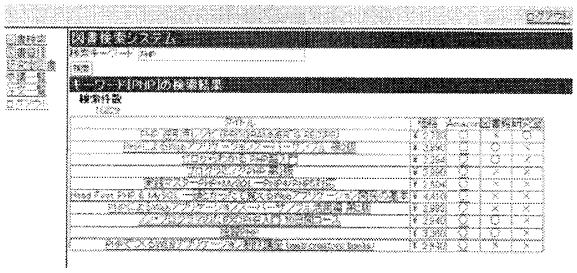

図 2 書籍情報の統括的検索

これまで述べてきた方法を検証するために、以下の情報源を利用するとして書籍情報の統括的検索の実現を試みる．

- **大学図書館蔵書検索(OPAC)**：大学生や教職員の図書に図書を貸し出すための蔵書検索サイト．人間向けの UI しか持たないとする．
- **Amazon ウェブサービス(AWS)**：Amazon の商品データを利用するためのウェブサービス．HTTP 経由の XML または SOAP、REST を介して利用する API が提供されている．

上記の情報源を統合して 2 種類の検索サービスを提供することを目標とする．

- **キーワード検索**：キーワード入力より検索を開始し、複数の情報源の統合した結果を表示する．図 2 に示すように検索結果より各情報源における同一書籍の状況を一覧できる
- **ISBN 確認検索**：一つ以上の ISBN を検索条件として入力し検索を開始する．検索結果に図書館に所蔵しない書籍のみ表示するので、書籍購入検討時に確認のために利用できる．図書館職員はまとめて確認を行う場合に便利である．

3.2 ラッパープログラムの実現

UI しか持たない OPAC 情報源を統合できるように、ラッパープログラムを開発する．OPAC の詳細検索画面の HTML ソースを分析し検索条件の入力と送信をプログラムから行えるように API を開発する．

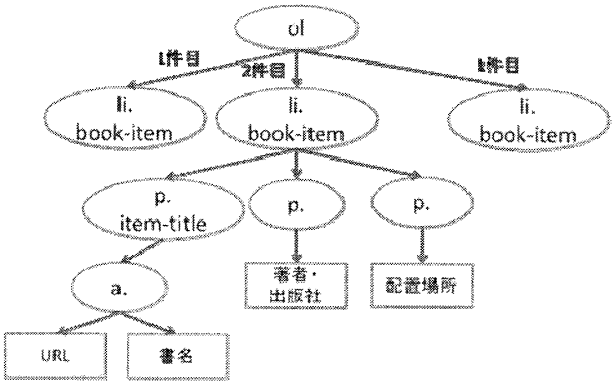

図 3 OPAC 検索結果の DOM 構造

検索結果を HTML ソースとして受け取り、その DOM 構造から 1 件ずつの書籍情報を調べる．図 3 では、大学図書館の検索結果の DOM 構造を示している．<ol>タグ以下に 1 件ずつの検索結果は"book-item"クラスの<li>タグで囲んでいるので、一つずつ調べていくことができる．また、書名は"item-title"クラスの<p>タグにあるので、そこから書名、詳細情報を表示するページの URL を取得できる．2 番目の<p>タグから著者・出版社の情報を得ることができる．

ここまで得られた情報は書名、著者名、出版社及び詳細情報へのリンクの URL しかなく、ISBN のような一意性のある情報はまだ入手できていない．必要に応じて詳細情報ページへのアクセスが必要になる．

4. 終わりに

本研究では我々は異種情報源統合の仕組みを考察したうえ、書籍情報の統括的検索の実現を試みた．ラッパープログラムの開発で UI しか持たない OPAC 検索を利用可能にした．

参考文献

[1]. G. Wiederhold and M. Genesereth. The Conceptual Basis for Mediation Services, IEEE Expert, Vol.12, No.5, pp.38-47 (1997).