単純な部分文字列照合による Web からの書誌情報の抽出

松本英樹
九州大学大学院システム情報科学府
〒 812-8581 福岡市東区箱崎 6-10-1
Tel: 092-642-3882
田中省作
立命館大学文学部
〒 603-8577 京都市北区等持院北町 56-1
Tel: 075-466-3301
池田大輔
九州大学附属図書館
〒 812-8581 福岡市東区箱崎 6-10-1
平木啓太
立命館大学文学部
〒 603-8577 京都市北区等持院北町 56-1

概要

現在，貸出履歴に基づく図書配架／推薦システムの構築を進めており，立ち上げ時のデータ不足や新刊情報の欠如などが問題となっている．本稿では，これらの情報を補完することを目的に単純な部分文字列照合による Web からの書誌情報の抽出のための調査について示す．

キーワード

書誌情報，表構造，書誌共起関係，Web, ライブラリ

Extracting Bibliographic Information from Web by Simple Substring Matching

Hideki MATSUMOTO
Graduated School of School of Information Science and Electrical Engineering, Kyushu University
6-10-1, Hakozaki, Higashiku, Fukuoka, 812-8581, JAPAN
Phone: +81-92-642-3882
Shosaku TANAKA
College of Letters, Ritsumeikan University
56-1, Toojiinkitamachi, Kitaku, Kyoto, 603-8577, JAPAN
Phone: +81-75-466-3301
Daisuke IKEDA
Kyushu University Library
6-10-1, Hakozaki, Higashiku, Fukuoka, 812-8581, JAPAN
Keita HIRAKI
College of Letters, Ritsumeikan University
56-1, Toojiinkitamachi, Kitaku, Kyoto, 603-8577, JAPAN

Abstract

There are a few issues concerning the sparseness of basic data for a bibliographical recommendation system, which is under development. These issues are primarily encountered at start-up or are related to new books. This paper shows the investigation for the extraction of bibliographical information from the Web by using simple substring matching for the comprehensive retrieval of such information for the system.

Keywords

Bibliographic Information, Table Structure, Cooccurrence Relation of Bibliographies, Web, Program Library

1. はじめに

現在，九州大学では学内のさまざまな分野の若手研究者による電子図書館プロジェクトを推進している．このプロジェクトは，従来のハード面を中心とした図書館業務の省力化・効率化や資料の電子化ではなく，情報技術を適用した，次世代の電子図書館としての新しいサービスとそのためのモデルの構築を目指したものである [2]．本稿では，プロジェクトの目指す成果の一つである，個人の図書貸出履歴に基づいた図書配架推薦システムと，そのための貸出履歴に代わる書誌情報を Web から抽出することについて考える．Web から得られる粗い書誌情報によって図書配架推薦システムの立ち上げ時や新刊に関する情報の欠如を補うことが期待される．なお，本稿の目的はあくまでも実装を念頭とした書誌情報の抽出で，技術的新規性よりもシステムのための実装，そしてライブラリ提供を優先することとした．

2. 図書配架推薦システムと現状の問題点

2.1 貸出履歴と書誌の共起情報

図書館における配架では，固定的に本が配列されているが，[3] のような仮想書架であれば，動的にさまざまな観点から図書を配架し，利用者に提示することができる．特に近年の自動書庫のような環境においては失われる，通常の開架での「本との出会い」を提供するものでもある．こういった本の連想関係のようなものを導く方法論は，データ・マイニングの典型的な応用として確立されつつある．その実装には，データを大量に抽出することが重要となる．その本と本のつながりを見出す最も基本的な情報となるものの一つが貸出履歴で，そこから「A という本を借りた人は B という本も借りている」という書誌の共起情報が得られる．また，このような情報は,「本と本」を結びつけるだけでなく,「本と人」「人と人」の結びつきを可能とするもので，多様な応用の可能性を有している．

2.2 問題点

図書配架推薦システム（以後，配架システム）で重要な役割を果たすのが，貸出履歴から得られる書誌の共起情報である．しかし，実際には次のような問題がある．

まず，個人情報に含まれる貸出履歴については，法的観点から現行の図書館では上記のような利用はできず，予め利用者にその利用承諾をとるなどの措置が必要である [1,2,3]．本稿では，この法的な議論と新しい運用モデルの詳細については割愛するが，今後，利用者が貸出履歴の提供の可否を選択し，このような情報の蓄積が進むことが期待される．詳細については [1,2] を参照されたい．

このような事情で，今後蓄積が進むことは期待されるものの，配架システムの立ち上げ時には，ほとんど貸出履歴はない．また，新刊書誌については当然，貸出履歴がない．こちらの問題については，利用可能な貸出履歴の量にかかわらず恒常的に起こるもので，より本質的な問題といってよい．

2.3 解決に向けての方向性

近年，Web 上には個人の読書履歴など，書誌に関する情報が掲載されている．そこで，このような Web 文書中の書誌情報を同定し，同文書中に含まれる書誌情報を，書誌の粗い共起情報と捉え，貸出履歴を補完するものとして収集する．配架システムは，貸し出し履歴が少ない書誌についてはこのような情報を拠り所とし，貸出履歴の蓄積が進むに従い，後者の情報を重用する．つまり，書誌の貸出履歴の蓄積量が進むにつれて，配架の計算から緩やかに Web 上の情報を排除すればよい．また，Web 上からは図書館に収蔵されている書誌とそれ以外の書誌，さらには収蔵されていない書誌同士の共起情報を得られ，配架システ

ム以外への応用も期待される（図１）．この収蔵されていない書誌を介した推移的な共起情報（図２）は，後述する Web 上での調査では収蔵されている書誌の 9 割程度を占めるため，非常に重要な役割を果たすものと考えられる．

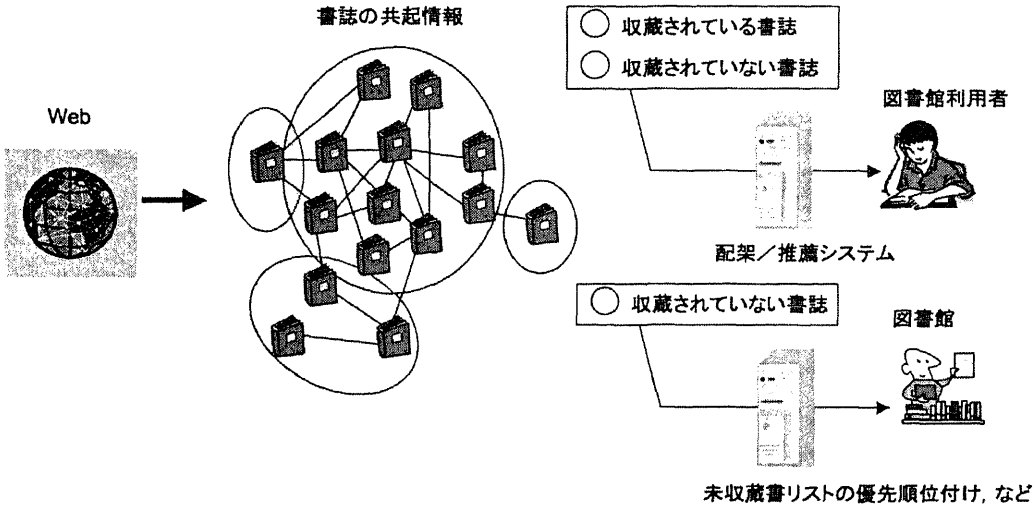

図１

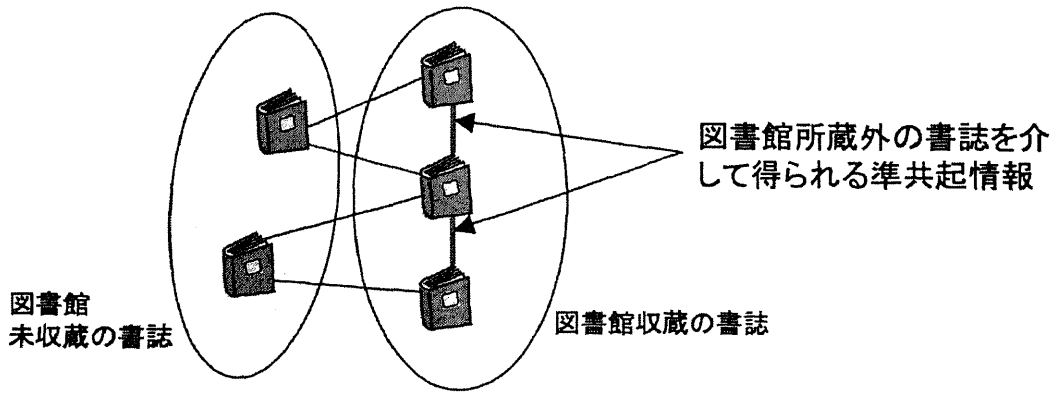

図２

3. 単純な照合法を用いた抽出の試み

3.1 ナイーヴな方法

本稿では書誌情報とは「書名」「著者名」「出版社名」の 3 つを考え，この 3 つの情報が同時に同定されれば，書誌が一意に（例えば，ISBN など）限定されるものとする．Web 文書内に記載されている書誌情報に関して，Web 文書作成者が読者に正確に書誌情報を伝達するという観点から，次のような素朴な仮定『あ

る程度の範囲内に曖昧さが生じない程度の書名の一部・著者名の一部・出版社名の一部が書かれている』を置く．

例えば,「書名：統計科学のフロンティア 10 言語と心理の統計ーことばと行動の確率モデルによる分析，著者名：岩波書店」を考えてみる．Web 文書中で書誌情報を書く際，書名におけるシリーズ名「統計科学〜10」や副題「ーことばと〜分析」は省略されることも多い．また著者名については，書名・出版社名に比べると複雑な構造をしており,「著」や「編」,「translated by」といった責任表示が加わり [4]，筆頭著者のみを示す場合（「甘利俊一他」）や苗字のみを列挙する場合（「甘利，金，村上，永田，大津，山西」）などさまざまである．出版社名については,「書店」や「出版」で終わる場合,「岩波」のように略記される場合がある．

議論の余地はあるが，仮定における「書名／著者名／出版社名の一部」を「書名／著者名／出版社名のプレフィックス」と考え，各情報で 3 つの情報で同定できる程度にプレフィックスが一致した場合，その書誌である可能性があると考えることにする（図３）．ただし，書名についてはシリーズ名や副題は書誌データベースで分離されていることが一般的であるので，シリーズ名・書名・副題を全て合わせた際のプレフィックスか，書名・副題のみの場合もある．また，複数照合するものがある場合は，一致するプレフィックスが長いものを優先する．

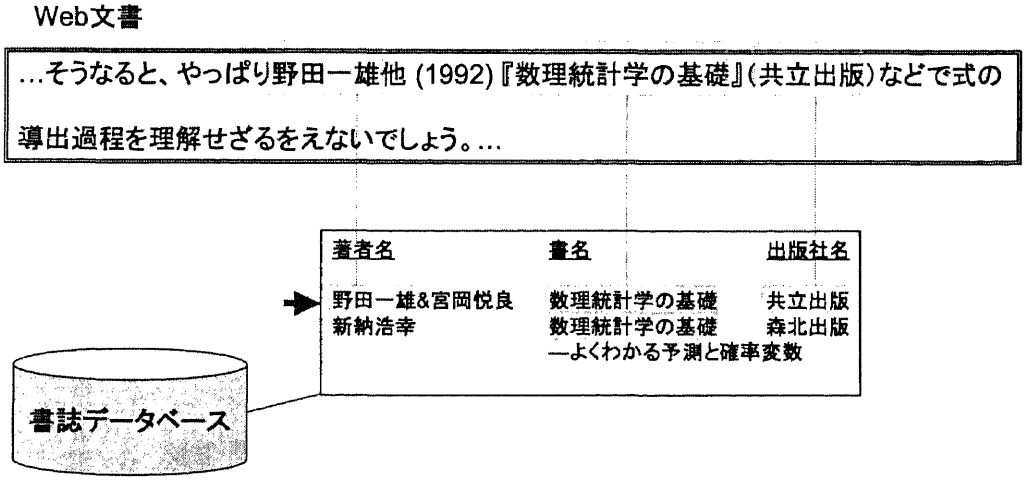

図３

予備調査として出版社名を検索エンジンへのクエリとして得られた Web 文書に対して，九州大学附属図書館（以後，九大図書館）の蔵書を管理する書誌データベースに基づき以上の処理を行った．例えば，一文中に一致するプレフィックスが書名 4 文字以上，著者名 3 文字以上，出版社名 2 文字以上が，書誌データベース内のデータと一致する場合に限定する．このようにして得られる書誌情報は全て正しいものとなるが，その数は極めて少ない．精度は高いものの，抽出数が少ないのには，主に次のようなことが考えられる．

（１）一般の Web ページに含まれる書誌情報は予想以上に少ない．

（２）九大図書館の書誌データベースには記載されていない，つまり収蔵されていない書誌も多い．

（３）Web ページに必ずしも書名・著者名・出版社名の情報を全て書いていない．

（４）クエリが書誌に依存したものではなく，よりきめ細かいクエリ生成が必要である．

また，このような方法では，書名や著者名がこの制限より短い場合は最初から抽出対象とならない，という問題がある．ここで（３）（４）は今回の検討の対象から外す.（２）は別の調査から，Web 上の書誌情報の 9 割近くは九大図書館の書誌データベースには記載がないものであることが推測されており，書誌データベースの制約から，非常に大きな情報を見逃していることになる．そこで，書誌データベースに記載されている書誌間の直接の共起情報だけでなく，収蔵されていない書誌についてもその共起情報を抽出ことを考える．書誌データベースには記載がない，というものも抽出するには，本節で述べた手法とは別の環境を仮定しなければならない．

3.2 表構造を考慮した方法

単純な部分文字列照合でも抽出数を犠牲にすれば，高い精度で一般の記述の中から書誌の同定は可能である．また，書誌データベースに記載がないものまで抽出するために，表構造に着目する．実際に Web ページを観察していると，表構造で書誌情報をまとめているものは数多く見られる．なお，抽出対象の表としては書誌情報のみが含まれているものを想定する．このような表構造を標的とすることで，書誌データベースに収蔵されている／されていないにかかわらず一括して書誌情報を抽出することが可能となる．したがって，問題は表が与えられたときに，書誌情報のみが記載された表か，そうではないのか，という判別問題に帰着される．

表構造については,『行または列はある性質で共通の要素が並んでいる』と仮定する．Web 上の表では多くの場合，行方向に各書誌の情報が並び，列方向に書誌の各情報が並ぶ．各セルに対して，3.1 の手法を適用することで，その列の属性を高い精度で判別することができる．表に対して次のような条件を課す（図４）．

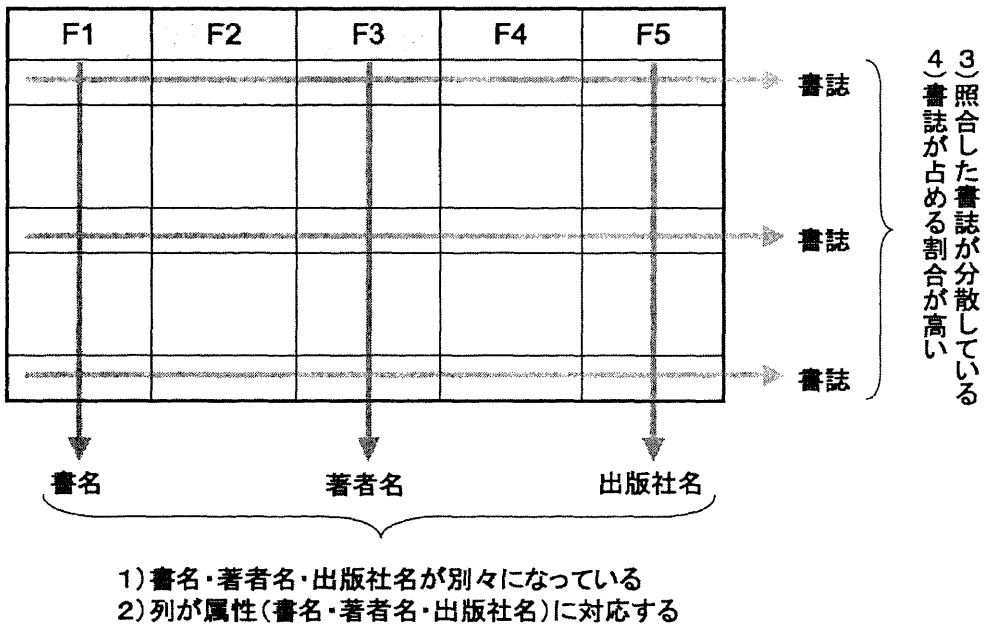

図４

（１）書名・著者名・出版社名に対応する列がそれぞれ独立に存在する．

（２）書名・著者名・出版社名に対応するデータから一意に書誌が同定される行が存在する．

を同時に満たし，

（３）同定された書誌が一様に分布している．

（４）表のサイズ（エントリ数）に対して書誌が占める割合が高い．

の少なくとも一方を満たすものとする．（３）はさまざまな定量化が考えられるが，ここでは行番号の相対位置に関する分散を考える．なお予備調査で検索エンジンに出版社名をクエリとして得られた文書 1494 ページに含まれる表に対して，（１）（２）だけでも 171 の表を全て正しく同定できるものとなっていた．なお，このデータで（３）の分散の平均は 0.34，（４）の平均は 5.8%である．ただし，（３）の分散の平均の計算において，一意に書誌が同定される行の存在しない表については対象外とした．これらの表から同定できた収蔵している書誌情報は 872 で，得られた共起情報は 2875 組となる．一方で，収蔵されていない書誌情報同士の共起情報は 2,386,325 組，収蔵されている書誌と収蔵されていない書誌の共起情報は 113,539 組得られる．

4. おわりに

本稿では，仮想書架などにおける配架システムに必要となる書誌の共起情報を Web から抽出する試みについて述べた．また，収蔵されていない書誌の存在が予想以上に大きいことが分かった．表構造を前提とすることによって，大量の書誌情報が得られる見込みは予備調査で分かったものの，そのほとんどは収蔵されていない書誌の共起情報である．したがって，このような共起情報を介して蔵書の共起情報が近似できるのかどうか，その相関はどの程度か，といった検討が必要である．一方で，本稿では割愛したが，書誌情報抽出について 3.1 の手法から得られる高品質の書誌情報からさらに抽出数を増やすための自然言語テンプレートの構築を試みたが，あまり良好な結果は得られていない．また，検索エンジンへのクエリの投げ方や，得られる情報を循環することなどにより，より大規模に抽出する枠組みを考える必要がある．ここで述べた手続きについては，現在ライブラリとして実装を進め，そのライブラリは公開する予定である．

データ・チェックを手伝って頂いた立命館大学 石川 徹氏に記して感謝の意を表する．本実装の一部には，「九州大学教育研究プログラム・研究拠点形成プロジェクト（D タイプ）」の助成による．

参考文献

[1] 安東菜穂子，池田大輔，田中省作：電子図書館と利用者のプライバシー―履歴・個人情報の保護と利用の両立を目指して―，ディジタル図書館，No.30，2005.

[2] 池田大輔，安東菜穂子，田中省作：ディジタルライブラリにおける履歴・個人情報の保護及び利用，ディジタル図書館，No.27，2004.

[3] 池田大輔：新たな電子図書館モデル構築に向けて，九州大学附属図書館「図書館情報」，Vol.40，No.3，pp.49-50，2005.

[4] 国立国会図書館，図書館員のページ：書誌データの作成及び提供，日本目録規則 1987 年版改訂版採用方針．http://www.ndl.go.jp/jp/library/data/ncr87_rev.html#7

[5] 宮川拓也，山口恭平，大森洋一，池田大輔：Web 上における仮想書架の試作と評価，ディジタル図書館，No.28，2005. http://takuya.ale.csce.kyushu-u.ac.jp/exp2/

[6] 野口正人，廣川佐千男：Web からの同系統単語知識獲得方式，2003 年情報学シンポジウム講演論文集，p.21-24，2003.