社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2007－DD－63 (5)
2007／9／28

区分記号の配置を利用しての書誌要素の自動抽出

林 典門、海尻 賢二
信州大学 総合工学系研究科

書誌情報はディジタルライブラリの基本情報として整備が進められている。書誌目録は各図書館によって様式が異なっており総合目録の作成が求められている。書誌要素の識別のためにタグを付しているが、このタグは書誌データにより全て異なっておりデータベース間のデータの相互照合、横断検索等の壁となっている。

このタグを意識することなく書誌事項の識別を自動的に行う方法を提案する。このためにSVMの考え方とHMMを利用して、区分記号と書誌要素の配列から書誌要素を判定して書誌要素の自動抽出をすることを研究目的としている。

Automatic extraction of the bibliographic elements using arrangement of delimiters

Norikado Hayashi、 Kenji Kaijiri
Shinshu University

Catalog information is different in a style by each library, and standardization of this information is called for. In the catalog, the bibliographic element is identified with the tag, but the usage of tag is different in each catalog. The contents of the bibliographic catalog and a bibliographic element can be identified from the arrangement of delimiters. We propose the identification method by using SVM and HMM.

1、まえがき

書誌情報のデータベースは多種多様であり多く目録情報の検索がネット上で検索ができる。国公立の図書館、資料を扱う機関ではこれまで蓄積してきた書誌情報を整理して総合目録を作成し、更に全国ネット用の総合目録データベースを作る構想が始っている。ディジタル図書館の基本となる書誌データベースは文献情報の検索の情報源としてその整備が求められている。これまで各図書館ごとに進められてきた書誌情報の電子化の構想を共通の規則でまとめて、何処の機関でも同じように利用ができることを目指している。しかし、実際の目録情報は図書館毎に様式が異なり目録情報の仕様の“標準化”が求められている。

書誌目録のデータベースには書誌要素ごとにタグを付して書誌要素で識別をする（タイトル、著者、価格、等）方式を採っている。このタグはデータベースにより記号が異っている、このためデータベース間のデータの照合、横断

検索等に不都合を生じている。

書誌要素は文字、数字、記号が混在した文字列である。この文字列を意味のある文字列に識別するために区分記号が付されている。この区分記号の種類、配置を分析して書誌要素を自動的に抽出するシステムを考える。

関連研究としては[1,2]の「ＳＶＭ／HMM」の技術を利用したOCRで読み取り文献データから書誌要素を自動抽出する研究がある。 一方本研究ではＷＥＢ上の書誌データベースからデータを読み込んで書誌要素を抽出する方法を研究する。

以下、２章では、その概要、３章では区分記号によるデータの抽出の方法、そして４章では実験と評価を述べる。

２、 基本概念

２-1 研究の位置付け

この研究の目的は、ＷＥＢ上の書誌目録をタグを頼らずに目録の書誌要素の前後に付く区分記号を判定することによって自道的に抽出して書誌項目（タイトル、著者、等）を識別する方法の研究である。多くの書誌データベースがＷＥＢ上に作られてお り、書誌項目はすべて共通しているが個々に別々の様式の使用タグ、書誌事項の配置の違いがある。これによりデータベース間の横断検索、共同利用に不都合も生じている。区分記号は目録規則により種類と配置に規則があり書誌事項の識別を可能にしている。これによりタグを頼らないで書誌目録から書誌要素を自動抽出することも可能である。こうした区分記号の役割を利用してあるデータベースのデータを抽出してそのデータを他のデータベースへ再配置をしたり、新しいデータベースに作り直すこともできる。この再配置の応用例として総合目録の作成、ダブリンコアのデータの自動作成等がある。目録情報の自動作成、再編成へ可能にすることにより書誌情報の応用分野を広げることになる。

２-2 研究方法の概要

抽出については以下の概念を利用する：

ＳＶＭ（ Support Vector Machine)：データを分析して適合する区分域にデータを分類をする機能で、区分記号の配置により抽出されたデータを識別する [2]。

HMMの基本 (Hidden Markov Model)：隠れマルコフの概念をもとにしての確率モデルを扱うシステムの構想で出現する過程が未知のマルコフ過程と仮定して、予測可能な情報からその未知の過程を推定する。この考え方 は動的計画法の一種であり、最尤状態の遷移列を示すものである。 ある時点 t での最尤状態遷移列は t までに観測された情報とすると、t-1 までで尤も確からしい最尤状態遷移列だけに依存するという仮定によるもので、t-2 以前のモデルパラメータには全く依存しないというアルゴリズムである。

３、書誌要素の自動抽出

本章では具体的な書誌要素の抽出の手法について述べる。大まかには１）区分記号による書誌候補の識別、２）SVM を利用した書誌要素の分類、３）隠れマルコフモデルによる書誌要素としての予測、の２段階のステップからなる。

３-1、 区分記号による方法とは

目録要素を自動抽出するためには目録情報の書誌要素の区分記号の配置により書誌要素の内容と役割と意味を指示している。この区分記号の識別により抽出されたデータの内容を分類することができる。書誌目録の書誌要素は目録規則の定めるところであるが、区分記号の役割、配置は目録の作成の要件により異なるこ

ともある。サンプルとして情報学研究所の入力マニュアルの記入方法を示す。利用する区分記号の種類は目録規則に従っており、その例は以下のようである[4]。

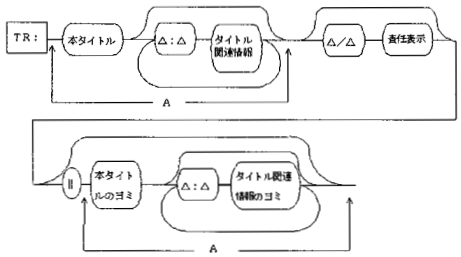
ピリオド、スペース（. △）

スペース、イコール、スペース（△ = △）

スペース、ピリオド、スペース（△ . △）

コンマ、スペース（, △）

スペース、プラス、スペース（△ + △）



3-2、 ＳＶＭの方法とは

書誌目録は書誌要素毎に区分記号が配置されている。この配置された区分記号を利用して記号間のデータを抽出することができる。多種類の区分記号を効率良く見分けてデータを抽出する。そのためには、書誌要素に使用する区分記号の識別と役割を定めておくことが必要である。文字列の分析と種々の区分記号の識別をプログラムによって見分けて処理して書誌要素の抽出をする。抽出されたデータの“正しさ”は文字列の状態、書誌要素の配置位置、出現回数、データの文字列の型、目録記述の内容（タイトル、著者、書誌事項等）の配列のパターンによる分類等で評価する。抽出された書誌要素の識別を更に正確にするための方法を具体化するものがＳＶＭの機能である。目録毎に抽出した書誌要素にナンバーを付ける。A1、A2…A9 として、次の目録の書誌要素には B1、B2、・・・、B9 として、全ての書誌要素を集合として表現する。さらに｛Ａｉ｝、｛Ｂｉ｝、・・・（ｉ＝1、・・・ｎ）として、タイトル、著者、等の書誌要素の要素別の集合を作る。これをタイトルはＳ１｛A1、B1、C1、・・｝、著者はＳ２｛A2、B2、C2、・・｝、・・と表現する。Ｓ１、Ｓ２、・・、Ｓｉ（ｉ＝1、・・・ｎ） は集合ナンバーとして、抽出した書誌要素の集合に付ける。この段階では、抽出された書誌要素は仮の書誌分類である。この分類集合を以下のように表す。

Ｓ１＝｛Ａ１、Ｂ１、Ｃ１、・・・　｝（タイトル）

Ｓ２＝｛Ａ２、Ｂ２、Ｃ２、・・・　｝（著者）

Ｓ３＝｛Ａ３、Ｂ３、Ｃ３、・・・　｝（出版地、）

・・・・・・

Ｓｎ＝｛Ａｎ、Ｂｎ、Ｃｎ、・・・　｝

次の工程として、抽出書誌要素と規定の分類の適合試行を行う。すなわち集合の書誌要素がタイトル、著者、等の書誌分類の何に適合するかの試行して分類をする。

Ｓ１～Ｓｎの書誌要素集合に確定した分類が付けられる。書誌要素の集合に書誌分類が付与される。（ＴＴ，ＡＡ、ＰＰ、・・・） この処理手続きとして区分記号の認識と使用の分析、規定・文法（区分記法）による考え方、（たとえば マルコフ連鎖、確率論、オートマトンそしてダイナミックプログラミング、チュウニングマシン）が必要になる。

3-3、 HMMによるタグの識別方法

マルコフ過程に従って事象が遷移する状態、および、各状態における記号の出現確率の分布よって構成される確率モデルを、「隠れマルコフモデル」と言っている。外部から見えるのは記号の系列的配置だけであり、内部の状態遷移は直接見ることはできない。ところから隠れマルコフと呼ばれている。この隠れマルコフモデルを目録情報の区分記号の系列的配置だけで分析をしてタグに頼らないで書誌要素の種類・配列を識別することに応用をする。

書誌情報への応用的考え方として目録記述において目録項目の要素 t（タイトル、著者、・・・）における要素を x_t と書くことすると書誌要素の配列をマルコフ過程では

$p(x_t|x_{t-1},x_{t-2},\cdots,x_{t-m})$

と表示する[1]。また、直前の要素のみに依存する単純マルコフ過程は、$p(x_t|x_{t-1})$ として書誌目録の要素の配列を表すことになる。この確率を使っての考え方をいる、これには段階的なプロセスが必要である。確率変数の各々の値に対して、その起こりやすさを表すことである。すなわち、起こりやすさの確率の値を対応させると言える。具体的にはタイトルの次には著者が現れる、さもなければ第２のタイトルが現れる。これは区分記号の記号識別により予測して“決める”ことが可能である。この場合　確率変数は離散的な値として取られるが、それ以上の理解はできない。 抽出した文字列がタイトルか著者かは区分記号の前後の関係の分析が必要となる。目録の種類により色々なパターンがあるが、書誌要素の配列は遷移の確率によって出現する書誌要素を特定する。例えば、

タイトル（ＴＴ）の次には著者名（ＡＡ）が現れる、

次に、著者名（ＡＡ）の次に出版者（ＰＰ）が現れる

という具合である。これには段階的なプロセスが必要であり、確率の値が遷移の起こりやすさを表している。起こりやすさの確率の値を対応させることである。このようにして区分記号の識別により遷移予測をすることが可能である。抽出した文字列がタイトルか著者かは区分記号の前後の関係の分析が必要となる。この分析にはマルコフ過程に従って事象が遷移する状態、および、各事象における区分記号の出現確率の分布よる確率モデル、等の「隠れマルコフモデル」を利用することである。

３･４、データへの適応

書誌目録データを材料としてその書誌要素を自動抽出する。区分記号の配置と抽出のための書誌事項タグのない目録、タグ使用の異なる種々の目録、特に区分記号の使用に規定のない目録、目録の要素配置に規定がない目録、等夫々の図書館の目録には異なる条件がある。書誌データベースの基本項目（書誌要素）は共通している[5]。

４、結論

各図書館の目録内容についての構成、区分記号、書誌要素の配列等の比較をした。

書誌要素の抽出には書誌要素であるタイトル、著者、書誌事項等の体系化されている文字列を区分記号により区分して配列している。この文字列を区分記号よる選択でデータを抽出してデータとして使用する。このために区分記号の認識と使用の分析　規定・文法、規定文法によるプログラムでの抽出の方法が必要であり、効果的なプログラムを作成する。データ抽出をするために、より詳細に分析をするプログラムが必要である。国会図書館の納本週報のデータから１１８件を抽出して文字列の分析を行った結果を表１に示す。区分記号による抽出の精度は完全なものではないが書誌要素の前後の区分記号を認識することにより抽出は可能である。

｛　区分記号　－　文字列　－　区分記号　｝の関係は書誌目録の文字列のパターンである。

このパターンを利用して文字列を抽出するため、大切な基本配列である。このパターンは書誌要素の任意の出現、繰り返しによって変化をするので区分記号は常に定位置にあるとは限らない、しかし｛区分記号―文字列―区分記号｝の関係により抽出された文字列の内容を識

別しなければならない。このためプログラムの作成については、文字列を挟む区分記号のすべての種類と配置の移動の関係も考慮しての抽出処理をする。区分記号の定位配置の関係は表２に示す通りである。

参考文献

[1]岡田崇、高須淳宏、安達淳　ＳＶＭ／ＨＭＭによる引用文献データの同定　情報処理学会、研究報告、2004-DD-43（11）

[2]高須淳宏、相原健郎　テキスト認識エラーモデルによる引用文献文字列からの書誌要素の抽出　電子情報通信学会論文誌　Ｄ－Ⅱ　Vol. J87－D－Ⅱ　NO. 6

[3]隠れマルコフモデル　：フリー百科事典『ウィキペディア（Wikipedia）

[4]目録システムコーディングマニュアル　国立情報学研究所編

[5]OPAC 説明用文書（検索）　H15 年度開発版 国立国会図書館

表１：国会図書館の納本週報のデータ（１１８件）から区分記号により抽出した書誌要素

大項目	中項目	件数	別項目	件数
タイトル（ＴＴ）		118		
	サブタイトル	61		
著者名（ＡＡ）		115	版表示	11
	訳者	11		
	編	11		
	監修	15		
出版社（ＰＰ）		118		
	発売社	8		
形態（ＰＳ）		117		
	DVD、カタログ	4		
標準番号（ＮＮ）	（JP 番号）	118		
	（ISBN）	114		
分類		118		
	細分類	25		

表２：区分記号の定位配置の関係

パターン①
本タイトル　△:△　タイトル関連情報　△／△　責任表示　△;△　2番目以降の役割の異なる責任表示
‖　本タイトルのヨミ　△:△　タイトル関連情報のヨミ　△＝△　並列タイトル　？　並列タイトルのヨミ
△:△　並列タイトル関連情報　△:△　並列タイトル関連情報のヨミ　△＝△／△　並列責任表示　‖
本タイトルのヨミ　△:△　タイトル関連情報のヨミ　？　（“‖”は“—”に置き換わる。“？”は不明）
パターン②
版表示　△／△　版の責任表示　△:△　2番目以降の役割の異なる責任表示　△＝△　並列版表示
△／△　並列責任表示　. △　付加的版表示　△／△　付加的版表示　の責任表示
パターン③
初号の巻次　△(　)　次号の年月次　終号の巻次
△(　)　終号の年月次　△;△　初号の巻次
△(　)　次号の年月次△＝△　終号の巻次
△(　)　終号の年月次
パターン④
出版地　△:△　出版者・　△、△　出版・頒布等
(　製作地等　△:△　製作者等　△、△　製作等の日付　)
パターン⑤
数量　△:△　その他の　△;△　大きさ　△＋△　付属資料
タイトルの種類　：　タイトル　‖　タイトルのヨミ
主記入フラグ表　：　件名　△--△　細目　‖　件名のヨミ　△--△　細目のヨミ　//　件名の種類