機械学習手法による著作の自動同定

宮田洋輔（慶應義塾大学） miyayo@slis.keio.ac.jp

1. FRBR とその影響

1997 年に国際図書館連盟（IFLA）が発表した『書誌レコードの機能要件』[1]（*Functional Requirements for Bibliographic Records*, FRBR）は図書館目録の概念モデルを提供した。FRBR の登場後，新しい国際目録原則の策定や AACR2 の改訂となる『資料の記述とアクセス（RDA）』の出版等，図書館目録の世界に大きな影響を与えている。

物理的な媒体に固定された資料を中心として考えられていた従来の組織化手法に対して，FRBR モデルでは書誌レコードの対象となる実体を第 1 グループの実体として，「著作」，「表現形」，「体現形」，「個別資料」の 4 つの実体からなる階層的な実体として捉えている。個別資料は手元の 1 つ 1 つの資料，体現形は具体的に本として表れているものであるが，「個別の知的・芸術的創造」[1]である著作と「著作の知的・芸術的表現」[1]である表現形は物理的な形態をもたない抽象的な実体として定義されている。このようなモデルの構築によって書誌記述のための論理的な枠組みを確立し，かつ利用者に対しても柔軟な集中機能の提供を可能にする。また，これらの実体間や，著者や団体のような，典拠レコードの対象となる実体群，「概念」や「場所」のような著作の主題となる実体群との関連を設定し，図書館目録におけるナビゲーション機能の向上も図られている。

このようにこれまでの資料組織法と異なるアプローチを持つ FRBR モデルに基づく形に再構築する必要がある。この FRBR モデルに基づく図書館目録の再構築を「FRBR 化 FRBRize」といい，FRBR 化は現在の図書館目録研究の主要な関心となっている。

2. FRBR 化の先行研究

これまでに作成されてきた書誌データを含む完全な FRBR 化のために，コンピュータを用いた機械的なアプローチと人手によるアプローチとの 2 つが考えられる。表 1 に両アプローチの概要を示した。

表 1 FRBR 化アプローチの比較

	機械	人手
研究事例	● OCLC Work-set Algorithm（2002） ● 谷口（2009）など	● Bates がシナリオを提示 （2003） ● 宮田（2009）
利点	● 作業コスト低 ● アルゴリズムに基づいた一貫性	● 共同分担作業 ● 機械では漏れる事例に対応できる可能性
欠点	● 書誌レコードの品質に依存	● 作業コスト高 ● 人によって捉え方が異なる可能性

機械的な手法に基づく，FRBR 化の手法として，OCLC の Work-set Algorithm[2]をはじめとする，著作キーの照合による手法がある。この方法では，機械可読式の書誌レコードから，実体を識別するのに有効なフィールドを抽出し，それらを組み合わせて著作キーを構築する。そして，構築した著作キーの照合によって，著作のクラスタを作成する方法である。

OCLC は，WorldCat.org の FRBR 化や，FRBR 化された小説検索システム FictionFinder の提供など，さまざまなサービスに Work-set Algorithm を利用している。谷口は，JAPAN/MARC 形式のレコードを用いて，著作キーの照合に基づく方式での著作同定の実験を行った[3]。

機械的な手法は，作業コストが低くアルゴリズムによる一貫性の高さが利点といえる。しかし，これまでの方法では，書誌レコードのなかに含まれたデータだけしか利用することができず，性能は書誌レコードの品質に依存する部分が大きいと言える。

一方，人手による著作の同定は，これまで目録作成者が目録を作成してきたのと同じように，1 つ 1 つの資料を人手による精査に基づいてい

ずれの実体に属するのかを判定する，というアプローチである。Batesは議会図書館のタスクフォースに提出した報告書の中で，これまで図書館が共同分担目録作業によって総合目録を構築してきたように，さまざまな図書館が著作の同定に参加しFRBR化された目録構築をおこなうシナリオを提示した[4]。FRBR研究会による活動は，Batesの提示したシナリオの例と言える[5]。

人間が資料そのものや書誌レコードを見て判断する場合，機械処理では漏れてしまうような事例にも対応できる可能性がある。一方で，著作と表現形のような抽象的な実体に複数の人間が携わる環境においては，認識の一貫性の問題が考えられる。宮田は，目録作成者によるFRBR実体同定に関する実験をおこない，現在のFBBRの報告書における各実体と関連の定義だけでは十分な一貫性が得られないことを明らかにした[6]。

3. 機械学習手法を用いた著作同定実験

3.1. 本研究のアプローチ

図書館目録のFRBR化における「著作」の同定という課題は，特定の著作を想定した場合に，その著作に含まれるか否か，という2値の分類問題とも考えることができる。近年，分類問題の自動化には，機械学習に基づく手法が盛んに用いられている。

機械学習とは「あらかじめ分類済みの訓練用文書の集合を用いてカテゴリの性質を学習させること」[7]であり，学習によって分類器を構築し未知のデータに当てはめ予測する方法である。ラベル付きの学習用のデータからパターンを学習し，未知のデータに当てはめる予測することを，「教師付き学習」と呼ぶ。

機械学習手法を用いることによって，文字列の照合では同定が難しかったような事例に対して，パターンによって著作の同定ができる可能性が考えられる。そこで，本研究は，FRBR化の著作同定の機械的処理の高度化を目的として，機械学習手法を用いて既存の書誌レコードからの著作の同定を試みる。

3.2. 対象著作

実験には，人手によって著作同定に関するラベル付けがなされたデータが必要である。そこで，FRBR研究会[5]で作成されたデータを利用した。このデータは，J-BISCに収録されたレコードの中から，古典著作を対象として，特定の著作に対して網羅性があるように検索されたレコード群に対して，著作同定のための規則に基づいて，ある著作に含まれるかどうかの判定とその根拠と，著作の出現箇所にマークアップされたデータである。また，1つのレコードが複数の著作に属する場合には，それらすべての著作に対して同定が想定されている。

このデータから，「竹取物語」，紀貫之著「土佐日記」，吉田兼好著「徒然草」を実験の対象著作として用いた。それぞれの著作の中で，2005年までにJ-BISCに収録されたレコード集合を学習用集合，2005年から2009年までに収録されたレコード集合を判定用の集合として用いた。各著作の概要を表2に示した。

表2 実験に用いた著作の概要

		竹取物語	土佐日記	徒然草
学習	サイズ	331	321	817
	同定数	114	107	157
	%	34.4%	33.3%	19.2%
判定	サイズ	67	51	160
	同定数	12	7	15
	%	17.9%	13.7%	9.4%

3.3. 素性

JAPAN/MARCレコードに記録されたフィールドの中から，著作の成立と関係のある素性として，「書名」，「著者名」，「分類記号」，「件名標目」を抽出した。以下では，それぞれの素性の抽出方法について述べる。

上述の通り，実験に用いたデータは複数著作に属するレコードに対してはすべての著作に同定がなされることが想定されている。そのため，それぞれ複数の要素が抽出できる場合に，優先度を設定し1レコードに対して優先的な1つの著作を設定するのではなく，直積集合を作成し，複数の入力用データを作成することとした。以下では，要素の抽出と抽出されたデータへの処

理について記す。

1) 書名
25x フィールドのサブフィールド$A から取得した。また，書名フィールドなどに含まれている著作の出現箇所のマークアップは除去した。

2) 著者名
25x フィールドのサブフィールド$F と 751 フィールドのサブフィールド$B から取得した。751 フィールドにデータが入力されている場合は 25x$F を用いるのではなく 751$B のデータを用いた。25x$F の役割を示す「‖」以降の文字列の除去，751$B の「‖」と「()」内の文字列を除去する処理をおこなった。

3) 分類記号
677 フィールドのサブフィールド A から，「日本十進分類法」を取得した。

4) 件名標目
658 フィールドのサブフィールド B から「一般件名標目」を取得した。

3.4. 手法

ナイーブベイズとサポートベクターマシン（SVM）を機械学習手法として採用した。それぞれの手法は，統計解析環境 R を用いた。

3.4.1. ナイーブベイズ

ナイーブベイズ（NB）は，ベイズの定理を応用した分類手法である。比較的単純な手法ではあるが，スパムフィルタリングやテキスト分類に用いられている。実験には，e1071 パッケージ[8]の naiveBayes 関数を用いた。

3.4.2. サポートベクターマシン

サポートベクターマシン（SVM）は，カーネル法を用いた 2 値分類器である。近年テキストの自動分類などの多くの研究に用いられている手法である。実験には，kernlab パッケージ[9]の ksvm 関数を用いて実行した。カーネルには，ガウシアンカーネルを用いた。

3.5. 性能評価

性能評価は，著作同定の F 値，著作ではないものを同定する際の F 値。すなわち著作非同定の F 値，判定精度（A）の 3 つの尺度を用いた。F 値は精度（P）と再現率（R）の調和平均として定義されている。

$$P = \frac{\text{同定した正解件数}}{\text{（非）著作と同定した件数}}$$

$$R = \frac{\text{同定した正解件数}}{\text{全(非)著作件数}}$$

$$F = \frac{2 \times P \times R}{P + R}$$

$$A = \frac{\text{同定正解数} + \text{非同定正解数}}{\text{全同定数}}$$

4. 実験結果

4.1. 用いる素性の比較

各著作に対して，1)書名＋著者名（T+A），2)書名＋著者名+分類（T+A+C），3)書名+著者名＋分類+件名（T+A+C＋S）のように，用いる素性の組み合わせを変更させて実験をおこなった。

実験の結果を表 3 に示した。概ね用いる属性を増やすことで，著作同定に対する性能が向上

表 3 実験結果

		竹取物語			土佐日記			徒然草		
		同定 F 値	非同定 F 値	判定精度	同定 F 値	非同定 F 値	判定精度	同定 F 値	非同定 F 値	判定精度
NB	T+A	0.57	0.51	0.61	0.53	0.75	0.76	0.79	0.85	0.87
	T+A+C	0.70	0.74	0.78	0.48	0.68	0.71	0.80	0.88	0.89
	T+A+C+S	0.80	0.77	0.81	0.29	0.55	0.58	0.87	0.89	0.90
SVM	T+A	0.61	0.81	0.76	0.55	0.80	0.80	0.75	0.93	0.88
	T+A+C	0.60	0.79	0.74	0.53	0.81	0.80	0.81	0.91	0.90
	T+A+C+S	0.78	0.87	0.82	0.38	0.71	0.72	0.80	0.91	0.87

した。『竹取物語』と『徒然草』に関しては，T+A+C+S の際に，著作同定と非同定の両方で高い性能を示した。一方，『土佐日記』では，書名のみを用いた時が最も性能が高く，属性の追加が著作同定の性能を下げる傾向にあった。

4.2. 学習用集合サイズによる影響

機械学習を行う際には学習用集合の作成コストが課題となる。そこで，学習用集合に用いるレコード数を変更させて，学習用集合の大きさが同定性能に与える影響を検討した。

『徒然草』を対象に，素性として書名・著者名・分類・主題を，分類器としてナイーブベイズを用いて，学習用集合のレコード数を，50 件から学習用集合のレコード数まで 1 件刻みでレコードを無作為抽出し，著作同定に対する F 値の変化を実験によって検証した。その結果，学習用集合に 500 レコード周辺から性能の収束が見られた（図 1）。

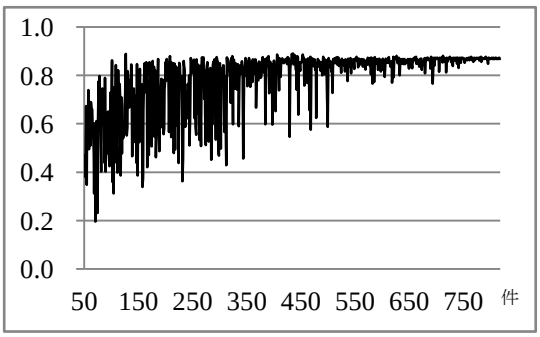

図 1 学習用集合サイズと同定 F 値の関係

5. 考察

機械学習手法を用いた著作同定の実験をおこなった。実験結果から，FRBR 化のための著作同定に機械学習の手法を用いることの有用性が示唆された。

本研究は，3 つの著作による小規模な実験環境で実験をおこなった。今後，他の著作群へ適用し，大規模な環境での性能の検証が考えられる。また，古典とは違う傾向を有すると予測される近代以降の著作への適用などが考えられる。性能向上のために，機械学習に用いる素性として，書名や注記中などに出現する語を用いた出現語の使用などが考えられる。

引用文献

1)書誌レコードの機能要件 : IFLA 書誌レコード機能要件研究グループ最終報告. 和中幹雄, 古川肇, 永田治樹訳. 日本図書館協会, 2004, 121p.

2)Hickey, Thomas B.; Toves, Jenny. FRBR Work-Set Algorithm. http://www .oclc .org/ research/ activities/ past/orprojects/frbralgorithm/2009-08.pdf, (2010-09-06).

3)谷口祥一. FRBR OPAC 構築に向けた著作の機械的同定法の検証: JAPAN/MARC 書誌レコードによる実験. Library and Information Science. 2009, no. 61, p. 119-151.

4)Bates, Marcia J. Task force Recom-mendation 2.3 Research and Design Review: Improving User Access to Library Catalog and Portal Information: Final Report(Version 3). Metadata Enrichment Task Force, 2003, 58p. http://www.loc.gov/catdir/bibcontrol/2.3BatesReport6-03.doc.pdf, (accessed 2010-09-06)

5)宮田洋輔. "著作同定における目録作成者間の一貫性". 三田図書館・情報学会研究大会発表論文集 2009 年度. 慶應義塾大学, 2009-9-26. 三田図書館・情報学会, 2009, p. 57-60.

6) フェルドマン, ローネン; サンガー, ジェイムズ. （辻井潤一監訳）テキストマイニングハンドブック. 東京, 東京電気通信大学出版局, 2010, 540p.

7)谷口祥一ら. "OPAC の FRBR 化を目指した人手による著作同定作業：FRBR 研究会の取り組み". 2010 年日本図書館情報学会春季研究集会発表要綱. 同志社大学, 2010-5-29. 日本図書館情報学会, 2010, p. 75-78.

8) Dimitriadou. Evgenia, et al. e1071: Misc Functions of the Department of Statistics (e1071), http://cran.r-project.org/web/packages/e1071/ , (accessed 2010-09-06)

9) Alexandros Karatzoglou, Alex Smola, Kurt Hornik, Achim Zeileis. kernlab : An S4 Package for Kernel Methods in R. Journal of Statistical Software. 2004, vol. 11, no. 9, p. 1-20. http://www.jstatsoft.org/v11/i09/, (accessed 2010-09-06)