機械学習を用いた図書館の資料選択に影響する要因の分析

安形 輝（亜細亜大学） agata@asia-u.ac.jp

1. 図書館における資料選択

図書館における資料選択は専門性の高い業務であると言われてきた。資料選択に関する研究は「価値論」「要求論」「目的論」に代表される抽象的な議論が中心である[1]。これらの研究の成果は図書館がどのような理論に基づき資料選択を行うべきかを示してくれる。しかし、これらの理論によって個々の資料を実際に選択できるわけではない。

より具体的な形で図書館における資料の選択や廃棄に関する研究が行われる場合、政治的立場や宗教などの点から論争的な資料を対象とすることが多い[2]。しかし、論争がない一般的な資料を具体的にどのように選択するかに関しては、研究ではなく実践的な話が雑誌記事等で言及されるのみである[3]。

図書館で資料選択がどのように行われるかを検討した研究の一つとして、長澤による「資料選択要因の考察」[4]が挙げられる。長澤は、資料選択に影響する要因として「図書館の目的」「利用者の要求」「コレクションの性格」「資料の性格」「スペースの問題」「財源の問題」「選択者の問題」を挙げている。ただし、どの要因がどの程度を与えるのかを実証的に明らかにしたものではない。

本研究では図書館の所蔵データ（つまり実際の資料選択の結果）に基づき、機械学習による分類器を用いて自動判別実験を行った。実験の目的は、1）図書館の資料選択に影響する要因を分析すること、2）資料選択の自動化の可能性を明らかにすることである。

2. 所蔵調査

所蔵の自動判別実験を行うためには、図書館が実際にどの資料を所蔵し、何を所蔵していないかを調べる必要がある。そのために、ある期間（2007 年一年間）の出版物を網羅するリストを作成し、そのリストに基づいて図書館 OPAC や検索 API を用いて所蔵調査を行った。

2.1 調査対象資料

2007年刊行図書の全点に対する所蔵調査となるように、以下の手順で対象を選定した。

(1) コミックを除く資料 （63,741 件）

2007 年度版『「Book」データベース』[5]収録中、ISBN が付与されており、日本図書コード（以下、C-Code。全 4 桁）の 3,4 桁目が“79”（コミックに付与するコード）以外の全ての図書 63,741 件とした。『「Book」データベース』はトーハン、日本出版販売、紀伊國屋書店、日外アソシエーツの 4 社が共同構築しているデータベースで、一般流通ルートに乗る図書をほぼ網羅している。ISBN を検索式とした調査を行うために ISBN が付与されたデータに対象を限定しているが、付与されていないデータはごく一部である（166 件）。

(2) コミック（8,209 件）

『「Book」データベース』には、コミックの多くが収録されていない。そのため、以下の手順で 2007 年に出版されたコミック 8,209 件の ISBN を取得した。1)Amazon[6]において、「漫画・アニメ・BL」カテゴリ、出版年が 2007 年である資料群を検索(9,181 件)。2)ISBN 誤付与資料を除去。3) c-code3,4 桁目が“79”である資料群を抽出した。

2.2 調査対象図書館

図書館の規模による違いを分析するために、市町村立図書館と都道府県立図書館の所蔵調査を行った。市町村立図書館と都道府県立図書館では調査手法が異なる。

(1) 市町村立図書館

市町村立図書館の所蔵調査は、図書館システム「CLIS/400」を採用している館を対象とした。理由は、以下の通りである。1)ISBN による検索が可能で「ISBN-10、ISBN-13」「ハイフン有無」の揺れに対応、2)他のシステムと比較し応答性能が良い、3)複本の調査も可能である。CLIS/400 の導入事例紹介ページ[7]によれば、全国 39 の自治体・団体が導入している。関東近県の 18 自治体（荒川区、葛飾区、墨田区、多摩市、中央区、豊島区、練馬区、世田谷区、西東京市、三鷹市、川崎市、調布市、新座市、三郷市、三芳町、東松山市、小平市）の 173 館について所蔵調査を行った。多くの自治体では自治体内での分担収集を行ない、

資料選択も自治体でまとめて行なっていることが多いと考え、資料選択実験では所蔵データを自治体単位で扱う。

(2)都道府県立図書館

都道府県立図書館については、国立国会図書館サーチの検索 API を用いて、総合目録サービス「ゆにかねっと」の登録館について所蔵を調査した(2012 年 9 月 10-16 日)。ただし、調査対象資料の所蔵が 1,000 冊に満たない、京都府立総合資料館、神戸市立中央図書館、神奈川県立図書館、神奈川県立川崎図書館、梅花女子大学図書館、三康図書館の 6 館は実験対象から除外した。複数館ある都道府県や政令指定都市の図書館をまとめることはせずに、64 図書館を対象とした。

実験対象とした自治体と図書館は合わせて 82 館となった。所蔵が多い上位 10 館は表1のとおりである。国立国会図書館は納本図書館として他の図書館とは異なり資料選択はおこなっていないが未納本がある。

表1 所蔵が多い上位10館(自治体含む)

図書館	所蔵冊数	カバー率
国立国会図書館	61,812	85.9%
大阪市立中央図書館	39,023	54.2%
岡山県立図書館	31,356	43.6%
さいたま市立中央図書館	30,149	41.9%
小平市立図書館(自)	28,421	39.5%
葛飾区立図書館(自)	26,648	37.0%
世田谷区立図書館(自)	24,825	34.5%
滋賀県立図書館	24,596	34.2%
川崎市立図書館(自)	23,699	32.9%
横浜市中央図書館	23,683	32.9%

3. 機械学習による資料選択実験

3.1 機械学習による分類器

分類器の実装としては Weka 3.7.7[8]を用いた。Waikato 大学(ニュージーランド)を中心に Java 言語で開発が行われているデータマイニングツールであり、数多くの機械学習に基づく分類器を実装している。多くの実装からここでは決定木とランダムフォレストを用いた。

(1) 決定木 C4.5

決定木(decision tree)は属性条件により分岐するノードから構成される木構造を用いた伝統的な機械学習手法の一つである。ここではWekaに実装されているC4.5[9](モジュール名は J48)を学習結果の分析のために用いた。決定木アルゴリズムの特徴は、与えられた属性に関する if-then ルールで木が構築されるため、他の手法と比べて可読性が高いことである。そのため、資料選択判定の可視化のために用いた。

(2)ランダムフォレスト(Random Forest)

多くの決定木を弱分類器として組み合わせて用いる集団学習に基づく分類器である。機械学習分野の分類器の中で、性能が高く、判定速度も高いと言われている。ここでは人手による資料選択をどこまで再現できるかを示すために用いた。

3.2 素性データの収集

図書館における資料選択には多くの要因が影響している。既往研究 [4]に基づき、できるだけ広く手がかりとなりうるデータを収集し、素性として分類器に投入した。

(1)書誌情報等:各資料の出版社、大きさ、ページ数、価格、ランキング、資料タイプの素性データは Amazon から取得した。タイトル、著者名は素性として用いなかった。

(2)C-Code:販売分類のコードであり、1 桁目が販売対象、2 桁目が形態、3,4 桁目が主題を表している。ジュンク堂サイト[10]より取得した。販売対象、形態、主題はそれぞれ異なる素性とした。

(3)選定図書総目録収録の有無:日本図書館協会作成の『選定図書総目録 2009 年版』[11]に収録された 2007 年の図書リストに基づき、選定図書総目録への収録の有無について情報を取得し、そのまま素性とした。

(4) 朝日新聞に掲載された書評:ブック・アサヒ・コム[12]に2007年に掲載された書評の507件についてISBNを取得した。書評として掲載されたかを素性とした。

3.3 評価尺度

この実験では精度、再現率、F 値を評価のために用いた。精度はどれだけ正確に所蔵を判定できたかを、再現率はどれだけ網羅的に所蔵を判定できたか、F 値は精度と再現率を組み合わせた総合的な尺度として用いた。

$$\text{精度} = \frac{\text{所蔵と判定された所蔵文献数}}{\text{所蔵と判定された文献数}}$$

表2 実験結果

図書館名	決定木C4.5			ランダムフォレスト		
	再現率	精度	F値	再現率	精度	F値
大阪市立中央図書館	0.875	0.804	0.838	0.860	0.814	0.837
岡山県立図書館	0.791	0.803	0.797	0.793	0.807	0.800
さいたま市立中央図書館	0.788	0.767	0.777	0.779	0.778	0.778
川崎市立図書館	0.705	0.766	0.734	0.713	0.771	0.741
荒川区立図書館	0.683	0.783	0.730	0.699	0.783	0.739
滋賀県立図書館	0.735	0.740	0.737	0.721	0.750	0.735
世田谷区立図書館	0.708	0.749	0.728	0.713	0.758	0.735
葛飾区立図書館	0.739	0.735	0.737	0.727	0.743	0.735
川崎市立中原図書館	0.695	0.756	0.724	0.704	0.762	0.732
大阪府立中央図書館	0.693	0.737	0.714	0.694	0.750	0.721

$$\text{再現率} = \frac{\text{所蔵と判定された所蔵文献数}}{\text{所蔵文献数}}$$

$$\text{F値} = \frac{2 \times \text{精度} \times \text{再現率}}{\text{精度} + \text{再現率}}$$

本実験では、学習用・判定用データを分割し、10 交差検定を行ったが、各データセットにおいて、各評価尺度の値を求め、それらを平均した値を算出した(macro-averaging)。

4. 実験結果

4.1 機械学習による資料選択性能

機械学習による資料選択の再現実験においてランダムフォレストを用いた場合、F 値は最高値が大阪市立図書館の 0.837、最低値が東松山市立図書館の 0.157、全 82 館平均値は 0.518 であった。ランダムフォレストにおいて F 値が高かった図書館上位 10 位を示したのが表2である。国立国会図書館に関してはF値が高かったが、この表からは除いた。

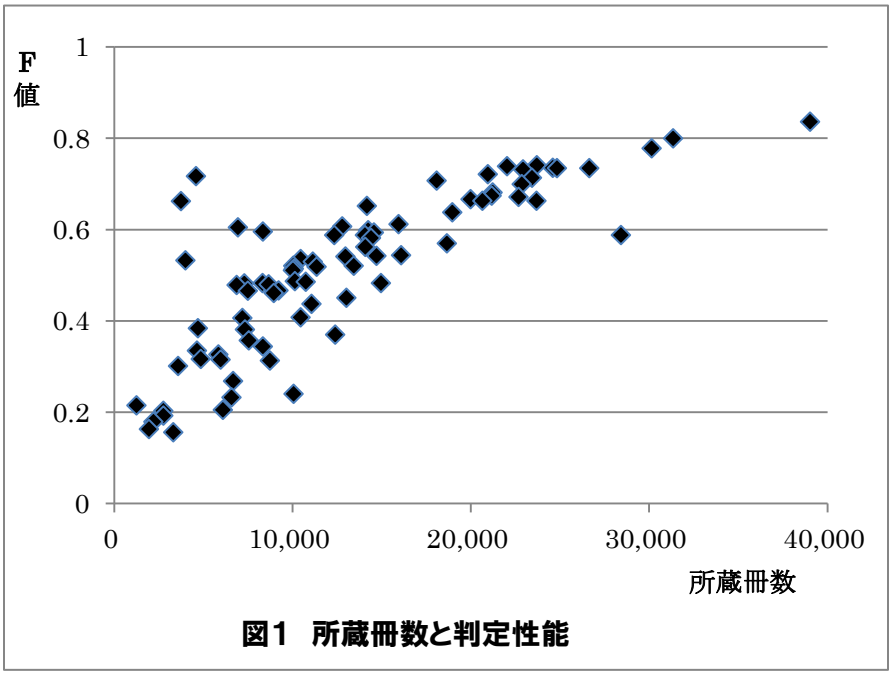

図1 所蔵冊数と判定性能

機械学習アルゴリズムの比較では一部の図書館を除き、ランダムフォレストがC4.5の性能を上回っており、F 値の平均は C4.5 が0.434、ランダムフォレストが 0.518 であった。最も高いF値を示したのは、C4.5 を大阪市立図書館

に適用した場合で、精度、再現率が共に高い。再現率は 87.5%であり、人が選択した資料をほぼ網羅している。

4.2 上位5館での判定に強く影響した素性

判定性能が高かった上位5館に関してはC4.5 などの機械学習アルゴリズムにおいて、素性選択に用いられている情報利得値を算出した。情報利得はある素性で分割する前の平均情報量と分割した後の平均情報量の差である。今回の実験においては所蔵をするかしないかにあたってその素性がどの程度影響するかを意味する。どの図書館に関しても「出版者」が最も情報利得値が高い素性であった。

4.3 所蔵冊数と資料選択性能

図1は各図書館の資料選択に関して、横軸に所蔵冊数、縦軸に F 値(全資料を対象としたランダムフォレスト)を取ってプロットした散布図である。ほぼ右肩上がりのグラフとなっている。所蔵冊数と F 値の間の相関係数を算出すると 0.81 であり、強い正の相関関係にある。所蔵冊数が多い図書館ほど、機械学習による資料選択の性能は高くなっている。散布図で左上にある外れ値の図書館は東京都立多摩図書館と大阪府立中央図書館国際児童文学館であった。この 2 館は所蔵資料数が少ない図書館の中では、機械学習による分類器が判別しやすい資料選択をしている。

4.4 コミックと資料選択

コミックは形態コード等から容易に識別できるため、コミックを除外した資料群からの資料選択の自動判別実験も行った。全資料群からの資料選択と比較した時に、F 値は全 82 館中 78 館の図書館で向上し、全体の平均値も 0.518 から 0.580 まで向上した。これはコミックの資料選択の難度の高さを示している。逆にコミックを除外した資料群からの性能が低くなったのは 4 館(葛飾区立図書館、小平市立図書館、横浜市中央図書館、滋賀県立図書館)であった。この 4 館はコミックの選択に関しては機械学習が判別しやすい方法をとっていると考えられる。例えば小平市立図書館はこの期間に刊行されたコミックの所蔵はしていない。

4.5 国会図書館の未納本の識別

国会図書館に関しては未納本(10,138 冊)を対象として自動判別を行った。精度は 0.805、再現率は 0.690、F 値は 0.742 であった。

5. まとめ

本研究では資料選択の自動判別実験を行った。資料購入数が多い図書館の資料選択に関しては自動化の可能性が高いことが明らかとなり、それらの図書館では、投入した素性のうち、資料選択に最も影響する要因は出版者であった。

【注・引用文献】

1) 資料選択の理論がどのように議論されてきたかについては、安井一徳. 図書館は本をどう選ぶか. 勁草書房, 2006, 169p.が詳しい。
2) 例えば「選択」に関しては、大場博幸. “所蔵における公平 : 公立図書館における「郵政民営化」または「靖国神社」を主題とする書籍の所蔵”, 常葉学園短期大学紀要, no.42, 2011, p.15-33、「廃棄」に関しては、前田稔. “思想の自由と「公的な場」の「公正」 : 船橋市西図書館蔵書廃棄事件判決の評価”. 図書館界, vol.58, no.3, 2006, p.154-163.
3) 例えば、座談会 書店さん図書館に言いたいこと言って! ず・ぼん, No.10, 2004, p.18-56.に選書会議や見計らいの現場に関する記述がある。
4) 長澤雅男. 資料選択要因の考察. Library Science, 1966, no.4, p.143-154
5) 「BOOK」データベース. http://www.nichigai.co.jp/dcs/index3.html
6) Product Advertising API. https://affiliate.amazon.co.jp/gp/advertising/api/detail/main.html
7) CLIS/400 導入館一覧. http://www.y-net.co.jp/clis/clis_user.html
8) Witten, Ian H.; Frank, Eibe. Data Mining: Practical machine learning tools and techniques, 3rd ed., San Francisco, Morgan Kaufmann, 2011, 629p.
9) Quinlan, J. R. AI によるデータ解析. 古川康一監訳. 東京, トッパン, 1995, 293p.
10) 丸善&ジュンク堂書店公式サイト. http://www.junkudo.co.jp/
11) 日本図書館協会編. 選定図書総目録 60 (2009 年版). 日本図書館協会, 2009. CD-ROM.
12) ブック・アサヒ・コム http://book.asahi.com/