

図書館の貸出履歴と書誌情報を用いた 図書推薦システムの有効性

辻慶太, 滝沢伸也, 佐藤翔, 池内有為, 池内淳, 芳鐘冬樹, 逸村裕

Effectiveness of Using Library Loan Records and Bibliographic Information for Book Recommendation, by TSUJI Keita, TAKI-ZAWA Nobuya, SATO Sho, IKEUCHI Ui, IKEUCHI Atsushi, YOSHIKANE Fuyuki and ITSUMURA Hiroshi.

本研究では(1) T 大学図書館の貸出履歴約 6 年分を用いたアソシエーションルールの確信度・支持度, (2) 図書のタイトル類似度, (3) NDC, (4) 「BOOK」データベースの要旨の類似度, の各種組合せを学習用データとしたサポートベクターマシン(SVM)を用いて図書推薦の実験を行った。さらに SVM を用いず上記情報を単独で用いた場合や Amazon による推薦も比較検証した。被験者は32名で, 結果, 「NDC+タイトル+貸出履歴」と「タイトル+貸出履歴」の組合せが有効であること, Amazon の推薦の方が評価が高いことを確認した。

1. はじめに

近年図書館界では, 利用者の貸出履歴を有効に活用する方法が議論されている。¹⁾²⁾³⁾⁴⁾⁵⁾ その中の一つに貸出履歴を用いた図書推薦システムの実現がある。ここで貸出履歴とは, 各利用者が借りた資料の ID, 書誌事項, 借りた年月日から構成される利用者ごとの貸出記録の総体を指し, 図書推薦とは図書のタイトル, 著者名, 出版者名, 刊年といった書誌事項を提示して, その図書の閲覧・貸出を利用者に薦めることを指す。貸出履歴を用いて図書推薦を行う研究としては原田, 原田・増田, Tsuji ら, 辻らがある。⁶⁾⁷⁾⁸⁾⁹⁾ だが彼らの手法が用いる情報は主に貸出履歴のみで, 各図書のタイトルは用いておらず, NDC (日本十進分類法) についてはヒューリスティックによる扱いにとどまっている。辻らが指摘するように, 貸出履歴は図書館の全蔵書に比すると少なく,

貸出履歴だけでは適切な推薦が行えない図書も多数生じる(例えば一度も借りられていない新着図書は貸出履歴だけでは適切な推薦が難しい)。そこで本研究では, 図書のタイトルや NDC, さらには「BOOK」データベース(以下 BOOKDB)¹⁰⁾ に収載されている各図書の要旨も, 貸出履歴と同様, 図書推薦に有効な手掛かりであると仮定し, それらの最適な組合せ及び重み付けを機械学習で決定し, 推薦を行う手法を提案する。具体的には, 利用者が「現在興味ある 1 冊」として挙げた図書に対して, 機械学習, 特に自動分類の代表的手法であるサポートベクターマシン(以下 SVM)を用いて, 貸出履歴, タイトルの類似度, NDC の一致不一致, BOOKDB の要旨の類似度, を学習用データとして組み合わせ, その中で図書推薦のパフォーマンスが最も高い手法を特定し提案する。貸出履歴については, 原田・増田の協調フィルタリングを用いた手法よりも辻らのアソシエーションルールを用いた手法の方がパフォーマンスが高かったことを受け, アソシエーションルールの確信度と支持度を SVM の学習用データとする。

さて辻らは図書館外部の図書推薦システムとして Amazon のそれを挙げ,¹¹⁾¹²⁾ アソシエーションルールによる推薦とパフォーマンスを比較し, Amazon の方が結果が良いことを見出している。従って図書

2013年9月2日受理

つじ けいた 筑波大学図書館情報メディア系
たきざわ のぶや 筑波大学情報学群知識情報・図書館学類
さとう しょう 同志社大学社会学部教育文化学科
いけうち うい 筑波大学大学院図書館情報メディア研究科
いけうち あつし 筑波大学図書館情報メディア系
よしかね ふゆき 筑波大学図書館情報メディア系
いつむら ひろし 筑波大学図書館情報メディア系

館の蔵書のうち利用者が興味を持った1冊(例えば借りたり OPAC でクリックした1冊)に対して, Amazon が「その図書を購入した人は以下の図書も購入しています」と提示する図書を Amazon の API¹³⁾で調べ, それらの中から図書館が所蔵する図書を表示すれば, 低コストでパフォーマンスの高い図書推薦システムを図書館で実現することが出来る。

以上の問題関心から本研究では, 後述する T 大学の学生32名を被験者として, 彼らに現在興味ある図書を1冊ずつ挙げてもらい, Amazon との比較も含めた図書推薦の実験を行った。実験ではまず以下の4種類の組合せを学習用データとした SVM で推薦を行った。即ち, (1) 確信度・支持度・タイトル類似度・NDC の一致不一致・BOOKDB 要旨の類似度, (2) 確信度・支持度・タイトル類似度・NDC の一致不一致, (3) 確信度・支持度・タイトル類似度, (4) タイトル類似度・NDC の一致不一致, の4種類である。さらに SVM は用いず, 単純に(5) タイトル類似度, (6) 確信度, (7) BOOKDB 要旨の類似度, がそれぞれ最も高い図書も推薦した。さらに先述のように, (8) 現在興味ある1冊に対して Amazon が推薦する図書のうち T 大学図書館が所蔵する図書も推薦した。以上, 本研究では8種類の推薦方法のパフォーマンスを比較検証した。

本論文は以下のように構成される。まず第2章では先行研究を紹介する。第3章では本研究で用いたデータについて述べ, 第4章, 第5章ではそれぞれ推薦手法, 評価手法について述べる。さらに第6章で結果を述べ, 第7章で考察を行った上で第8章で総括する。なお本研究は館種としては大学図書館を取り上げ, 実験・調査の対象とする。ここで得られた知見のいくつかは公共図書館においても有効と思われるが, その精査・検証は今後の課題としたい。

2. 先行研究

図書館における図書推薦に貸出履歴を用いた研究は少ない。先述のように協調フィルタリングを用いた研究としては原田, 原田・増田があり, 協調フィルタリングとアソシエーションルールを用いた研究としては Tsuji ら, 辻らの研究がある。¹⁴⁾¹⁵⁾¹⁶⁾¹⁷⁾ まず原田は貸出履歴として, ある大学図書館の利用者8,808人分の960,078件のデータを用いている。被験者は「公開できる貸出図書数が21冊以上ある」大学生35名で, 彼らに協調フィルタリングによって30冊

ずつの図書を推薦し評価してもらっている。次に原田・増田は貸出履歴として, ある大学図書館でこれまで200冊以上500冊未満の図書を借りた学生の貸出履歴272,751件を用いている。被験者は先ほど同様公開できる貸出図書数が21冊以上ある大学生10名で, 彼らに協調フィルタリングによって図書を推薦し評価してもらっている。先ほどの原田と異なる点は, 原田・増田は図書の NDC の類を重み付けに用い, 借りた年月によって貸出履歴を分割している点である。原田・増田はこうした重み付けや分割に関して5種類のパターンを試し, 最も評価が高いパターンを特定している。

Tsuiji ら, 辻らは T 大学図書館の利用者39,442人分の1,854,345件のデータを用いている。被験者は大学院生・学部生合わせて33名で, 彼らに協調フィルタリング, アソシエーションルール, Amazon による推薦図書を提示し, 評価してもらっている。結果, 被験者の評価は Amazon>アソシエーションルール>協調フィルタリングの順に有意に高いことを見出している。

他に貸出履歴を用いた図書推薦研究としては Whitney & Schiff がある。¹⁸⁾ Whitney & Schiff は U CLA (University of California, Los Angeles) 図書館の貸出トランザクションをデータとして用い, 手法としてはグラフモデルに基づく推薦手法を提案している。彼らの手法は本研究が取り上げるアソシエーションルールと本質的に近い。

実験を行わずに図書推薦の手法を提案した文献としては, Shirgaonkar ら, Chen & Chen, Luo らなどがある。¹⁹⁾²⁰⁾²¹⁾ また推薦アルゴリズムの解説としては神鷹,²²⁾²³⁾²⁴⁾ Jannach ら²⁵⁾ が詳しい。

3. 用いたデータ

3.1 貸出履歴と蔵書

本研究では, T 大学図書館の貸出履歴をもとに図書の推薦実験を行う。²⁶⁾ 貸出履歴を構成する各貸出記録は資料1タイトルの書誌情報とある年月日にその資料を借りた利用者1名の情報から成る。具体的にはその資料の資料番号, 著者名, タイトル, 請求番号といった情報と, その図書を借りた利用者の属性情報から成る。上記貸出記録は2006年1月2日から2012年3月31日までのものであり, 全部で2,324,418件あった。このうち学部生によるものは999,630件, 大学院生・教職員によるものは1,294,012件, その

November 2013

他利用者(卒業生・一般学外者・名誉教授など)によるものは30,776件であった。本研究では「その他利用者」による貸出記録は用いず、学部生・大学院生・教職員による貸出記録2,293,642件(=999,630件+1,294,012件)を用いることにした。²⁷⁾これらの異なり図書数は477,668冊,異なり利用者数は44,571人(うち学部生は21,279人)であった。

T 大学図書館の貸出記録は先述のような形式を取っていることから、1人の利用者が同時に借りた資料の組(いわゆる1バスケット中のアイテム)は分からない。そこで本研究では同じ利用者が同じ年月日に借りた資料群は同時に借りられたとみなすことにした。学部生・大学院生・教職員によるバスケット数は821,771であった。

本研究では T 大学図書館の蔵書のうち NDC が付与されている図書643,676冊を対象とし、ここから推薦図書を選んだ(即ち、NDC が付与されていない図書は推薦対象としなかった)。洋書には NDC が付与されていないものが多い為、この冊数は T 大学図書館の全蔵書数よりも少ない。

3.2 「BOOK」データベース

「BOOK」データベースは(株)トーハン、日本出版販売(株)、(株)紀伊國屋書店、日外アソシエーツ(株)の4社で共同構築している図書情報データベースである。²⁸⁾本研究では2005~2011年度版を用い、それらに記載されていた図書の要旨(いわゆる帯やジャケットの掲載文)を図書推薦の情報とした。上記の各年度版において、要旨が添えられていた図書の数はそれぞれ44,143, 45,327, 48,241, 48,214, 48,459, 45,654, 44,652であった。平均して約75%の図書に要旨が添えられていた。

3.3 現在興味ある1冊

本研究では後述する32名の被験者に、現在 T 大学図書館が所蔵する図書の中から、研究や学習で用いる為に借りたい図書を1冊挙げてもらった。以下ではこれを「現在興味ある1冊」と呼ぶ。推薦する図書の決定は、この図書とのタイトル類似度や BOOKDB の要旨の類似度、NDC の一致不一致、アソシエーションルールによる確信度や支持度、などを用いて行った。

注ほか：図書館の貸出履歴と書誌情報を用いた図書推薦システムの有効性

3.4 学習用データ

SVM の学習用データとしては以下の2種類を用意した。

(1) まずまずは被験者33人に現在興味ある1冊を挙げてもらい、その図書に基づいて延べ460冊の図書を推薦し、興味の度合いを本研究と同じ枠組み(後述)で評価してもらっている。本研究ではこれらのうち「2:強い興味」「0:興味なし」と評価された図書をそれぞれ正例、負例とし、1つ目の学習用データとした。即ち、現在興味ある1冊と推薦した各図書のペアに対して、タイトル類似度、NDC の一致不一致、アソシエーションルールの確信度と支持度²⁹⁾、BOOKDB 要旨の類似度、を算出し、それらを4.3節で述べる x_i とした。そして「2:強い興味」「0:興味なし」をそれぞれ+1, -1とし、これらを4.3節の y_i として SVM の学習用データとした。

NDC が複数付与された図書に関しては、それぞれを独立した図書と捉えて学習用のペアを構成した。例えば現在興味ある1冊 A に対して2冊の図書 B, C が推薦され、A, B それぞれに2つの NDC が付与されていた場合、A は NDC ごとに A1, A2 という異なる2冊とし、B は同じく B1, B2 という異なる2冊とし、(A1, B1), (A1, B2), (A1, C), (A2, B1), (A2, B2), (A2, C), という6つの学習用ペアを作成した。上記のような方針のもと本研究では、186の学習用ペア(サンプル)を得た。正例、負例のペアの数はそれぞれ59, 127である。

(2) 2つ目の学習用データは以下の通りである。まず今回の被験者32人に対し、現在興味ある1冊に基づいて、(a)先ほどの学習用データを用いた SVM が出力した、正例への所属確率(後述)が最も高い6冊、(b)タイトル類似度が最も高い図書6冊、(c)アソシエーションルールの確信度が最も高い図書6冊、(d) BOOKDB 要旨の類似度が最も高い6冊、(e) Amazon が推薦する6冊、を提示し興味の度合いを評価してもらった。³⁰⁾³¹⁾そして被験者32人を16人ずつのグループ S と T に分け、グループ S が「2:強い興味」「0:興味なし」と評価した図書から上記と同様にして172の学習用ペアを作成した。これを先ほどの1つ目の学習用ペアと合わせ、合計358の2つ目の学習用ペア(サンプル)を作成した。正例、負例のペアの数はそれぞれ111, 247である。

以下では、2つ目の学習用データを用いた SVM が推薦した図書に対して、上記グループ T が下し

た評価を中心に述べていく。即ち、SVM に関しては16人の評価結果が中心となる。一方、上記(b)(c)(d)(e)のように学習を必要としない推薦については32人の評価結果を挙げる。以下の表で手法ごとにサンプル数が異なるのはこの為である。³²⁾なお、1つ目の学習用データに基づく結果も最後に述べる。

4. 推薦方法

4.1 タイトル・BOOKDB 要旨の類似度による推薦

被験者が現在興味ある1冊として挙げた図書とタイトルが類似した図書を推薦すれば、被験者から良好な評価が得られるかもしれない。そうした仮定の下、以下の手順で図書の推薦を行った。(1) 図書館の全蔵書のタイトルと被験者が現在興味ある1冊として挙げた図書のタイトルを共にMeCab ver. 0.994で単語に区切る。³³⁾(2) それら単語列から単名詞と2名詞列(以下ではこれらを「名詞列」と呼ぶ)を抽出する。³⁴⁾(3) 「全蔵書における出現タイトル数が5以上の名詞列」のTFIDFから成るベクトルを図書ごとに構成する。ここで名詞列 S の図書 A におけるTFIDFは「図書 A のタイトル中の S の出現頻度 $\times \log(\text{全蔵書数}/S \text{をタイトルに含む図書数})$ 」と定義する。(4) 被験者が現在興味ある1冊として挙げた図書 X のベクトルを (x_1, \dots, x_N) 、ある図書 Y のベクトルを (y_1, \dots, y_N) とすると(ここで x_i, y_i はそれぞれ対応する語の X, Y におけるTFIDF)、両図書のタイトル類似度は、本研究では以下のいわゆるコサイン尺度 sim で定義する：

$$\text{sim} = \frac{\sum_{i=1}^N x_i y_i}{\sqrt{\sum_{i=1}^N x_i^2} \sqrt{\sum_{i=1}^N y_i^2}}$$

(5) 全蔵書の中で sim が最も高い図書6冊を被験者に推薦する。(6) 上記タイトル類似度は後述のSVMによる推薦でも素性(=分類に用いる情報)として採用する。

上記(2)で2名詞列も用いるのは、単名詞だけよりも主題の特定性が高い名詞列が手掛かりとして使えるようになる為である。3名詞列以上を採用しないのは、3名詞列以上も採用すると異なり名詞列数が多くなり、計算時間が非常に大きくなる為である。(3)で出現タイトル数に下限を設け、全ての名詞列は採用しないのも、全ての名詞列を採用すると異なり名詞列数が多くなり、計算時間が非常に大きくな

る為である。こうした方針は実際にいくつかのパターンでタイトル類似度を算出し、値が高くなる図書とかかる時間を調べ、試行錯誤の末に決定した。ただしこれらは計算機環境に依存する面も大きいので、今後制限を緩めることも課題として残したい。

BOOKDBの要旨についても上記と同様に要旨同士の類似度を算出し、最も値が高い6冊を推薦した。またこの類似度は後述するSVMでの素性としても採用した。³⁵⁾

4.2 アソシエーションルールの確信度による推薦

本研究では辻らが用いたアソシエーションルールとほぼ同じものを採用した。³⁶⁾³⁷⁾³⁸⁾まず n 冊の図書 $X_i (i=1, \dots, n)$ をある利用者が一緒に借りた場合、その貸出履歴を $\{X_1, \dots, X_n\}$ と表し、これをトランザクションと呼ぶ。例えば図書 A, B, C を一緒に借りた場合のトランザクションは $\{A, B, C\}$ と表す。このトランザクションからは「図書 A を借りる利用者は図書 B も借りる」というルールを引き出すことが出来る。³⁹⁾従って、ある図書 X を借りる利用者は図書 Y も借りるというルールを $X \rightarrow Y$ と表すと、上記トランザクションからは $A \rightarrow B, A \rightarrow C, B \rightarrow A, B \rightarrow C, C \rightarrow A, C \rightarrow B$ という6個のルールを引き出すことが出来る。矢印の左側を前提部、右側を結論部と呼ぶ。

ルール $X \rightarrow Y$ の確かさ、有益さとしては確信度と支持度などがある。確信度は「 X, Y の両方を含むトランザクションの数/ X を含むトランザクションの数」、支持度は「 X, Y の両方を含むトランザクションの数/全トランザクションの数」である。本研究では、被験者が現在興味ある1冊として挙げた図書を前提部 X とするルールのうち、確信度が0.01以上、支持度が0.000001以上のルールを用い、これらの中で最も確信度が高い Y を6冊推薦した。⁴⁰⁾だが推薦においては確信度だけでなく支持度の高低も考慮した方が良いかもしれない。そこで次節で述べるSVMの素性には確信度と支持度の2つを用いた。⁴¹⁾

4.3 複数の情報を用いたSVMによる推薦

アソシエーションルールによる推薦では確信度だけでなく支持度も高い図書を推薦した方が有効と思われる。だが現実には両方とも高いという図書は少なく、一方のみ高いという図書が多い。ではどのよ

November 2013

うな値の組合せの図書を推薦するのが有効なのか？
 そうした問題には推薦に対する被験者の評価結果を
 コンピュータに学習させ、最適な値の組合せを特定
 するのが効率的である。即ち、被験者の評価結果を
 「良い」「悪い」の2つに分け、どのような値の組合
 せを「良い」に分類すべきかをコンピュータに学習
 させるのが効率的と思われる。さらに図書推薦では、
 確信度・支持度だけでなく、現在興味ある1冊との
 タイトル類似度や BOOKDB 要旨の類似度が高い
 図書を推薦した方が有効かもしれないし、NDC も
 考慮した方が有効かもしれない。そこで本研究では、
 そうした複数の情報を用いた自動分類の代表的な手
 法である SVM を用いた推薦も行った。採用した情
 報の組合せは以下の4通りである：⁴²⁾

- (1) タイトル類似度+確信度+支持度
- (2) NDC の一致不一致+タイトル類似度
- (3) NDC の一致不一致+タイトル類似度+確信
度+支持度
- (4) NDC の一致不一致+タイトル類似度+確信
度+支持度+BOOKDB 要旨の類似度

NDC の一致不一致では、NDC の類網目3桁全
 てが一致した場合は1、一致しなかった場合は0と
 した。⁴³⁾

以下では本研究が用いた SVM による推薦方法に
 ついて述べる。本研究では、SVM の代表的ソフト
 である LIBSVM ver. 3.12 を用いた。⁴⁴⁾ SVM には
 様々なバリエーションがあるが、最も一般的でかつ
 LIBSVM でもデフォルトになっている L1ソフト
 マージン SVM の“C-Support Vector Classification”
 (以下、C-SVC) と動径基底関数 (RBF: Radial Basis
 Function) カーネルの組合せを用いた。⁴⁵⁾⁴⁶⁾ 以下のパ
 ラメータ C と RBF カーネルのパラメータ γ の最
 適な値は easy.py で自動的に決定した。⁴⁷⁾ また各クラ
 スへの所属確率を出力させる為、-b オプションを
 1にした。被験者には、学習用データの「2：強い
 興味」というクラスへの所属確率が最も高い図書6
 冊を推薦した。

以下ではまず C-SVC について説明する。⁴⁸⁾ $\mathbf{x}_i (i=1, \dots, l)$ を各図書の素性の値から成るベクトルとする。
 例えば、被験者の現在興味ある1冊を A とした時、
 A と図書 i のタイトル類似度が0.3、アソシエ
 ションルールにおいて A を借りた人が図書 i も借

じはか：図書館の貸出履歴と書誌情報を用いた図書推薦システムの有効性

りするというルールの確信度が0.5、支持度が0.01で
 あったとする。この3つの値を使って C-SVC で学
 習を行う場合、 \mathbf{x}_i は例えば(0.3, 0.5, 0.01) など
 となる。また y_i は図書 i が学習用データにおいて
 「2：強い興味」と評価されていた場合は+1, 「0：
 興味なし」と評価されていた場合は-1を取る変数
 とする。この場合、以下の最適化問題として \mathbf{w} と
 b を求めることで SVM 分類器を構成することが出
 来る：

$$\min_{\mathbf{w}, b, \xi} \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^l \xi_i$$

$$\text{ただし } y_i(\mathbf{w}^T \phi(\mathbf{x}_i) + b) \geq 1 - \xi_i \quad \text{でかつ } \xi_i \geq 0 \\ (i=1, \dots, l) \quad \text{かつ } C > 0$$

ここで $\phi(\mathbf{x}_i)$ は \mathbf{x}_i を高次元にマッピングする関数
 で、後述のカーネル関数を構成するものである。 \mathbf{w}
 は係数(列)ベクトル、 b はバイアス項、 ξ_i はスラッ
 ク変数、 C はマージンパラメータである。

さて一般に \mathbf{w} 及び $\phi(\mathbf{x}_i)$ は非常に高次元になる
 為、上の問題を直接解くことは避け、以下の双対問
 題を解く方が多い：

$$\min_{\alpha} \frac{1}{2} \alpha^T Q \alpha - \mathbf{e}^T \alpha$$

$$\text{ただし } y^T \alpha = 0 \quad (0 \leq \alpha_i \leq C \quad (i=1, \dots, l))$$

ここで Q は $Q_{ij} \equiv y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$ の行列であり、 $K(\mathbf{x}_i, \mathbf{x}_j) \equiv \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$ はカーネル関数で、 $\mathbf{e} = (1, \dots, 1)^T$
 である。この最適化問題が解けた後は、 \mathbf{w} は以下
 のように算出できる：

$$\mathbf{w} = \sum_{i=1}^l y_i \alpha_i \phi(\mathbf{x}_i)$$

最終的に被験者には、以下の式の左辺の値がプラス
 になる \mathbf{x} を持つ図書を「2：強い興味」クラスに
 所属する図書として推薦する：

$$\text{sgn}(\mathbf{w}^T \phi(\mathbf{x}) + b) = \text{sgn} \left(\sum_{i=1}^l y_i \alpha_i K(\mathbf{x}_i, \mathbf{x}) + b \right)$$

$\mathbf{w}^T \phi(\mathbf{x}) + b = 0$ がいわゆる分離超平面であり、正例

負例間でマージンが最大になるものである。詳しくは阿部⁴⁹⁾などを参照されたい。

本研究では被験者の負担を考え、推薦する図書は1手法当たり6冊にとどめる。そこで「2：強い興味」に所属すると判定された図書の中から、その所属確率が最も高い図書6冊を被験者に提示する。LIBSVMでは学習時に**-b** オプションを1にすることでこの確率を算出させることが出来る。算出方法は以下の通りである。まずクラスの数 k 個あり、素性ベクトルが \mathbf{x} と表される図書 X が、クラス i に所属する確率を p_i と表す ($i=1, \dots, k$)。さらに k 個のクラスの中から i と j という2つのクラスを選び、図書 X がこれらのいずれかに所属するとしたら i の方に所属する確率を以下の r_{ij} で表す：

$$r_{ij} \approx P(y=i | y=i \text{ or } j, \mathbf{x})$$

この時、 $\hat{f} = \mathbf{w}^T \phi(\mathbf{x}) + b$ とすると、 r_{ij} は以下のように表すことが出来る：

$$r_{ij} \approx \frac{1}{1 + e^{A\hat{f} + B}}$$

ただし係数 A 、 B は学習用データの5分割交差検定法によって決定する。⁵⁰⁾ここでは k クラスに一般化して説明したが、本研究のように2クラスの場合は r_{ij} は p_i に等しい。

4.4 Amazon による推薦

被験者の「現在興味ある1冊」を Amazon で検索し、「この商品を買った人はこんな商品も買っています」と表示された6冊のうち T 大学図書館が所蔵しているものを OPAC で特定し、各被験者に推薦した。⁵¹⁾上記の検索及び特定は、本研究では手作業で行ったが、Amazon が提供している API を用いれば自動で行えるようになるはずである。⁵²⁾

5. 評価方法

以下ではまず本研究の被験者について述べ、次に被験者への推薦図書の提示方法と判定カテゴリーについて述べる。

5.1 被験者

本研究では T 大学で図書館情報学を専攻している学生32名に図書を推薦し、その適切さを判定して

もらった。⁵³⁾これら被験者32名の内訳は大学院生13名、学部4年生12名、学部2年生7名である(なお以下ではこれら3種類のグループは便宜上「学年」が異なると呼ぶことにする)。

5.2 被験者への提示方法と判定カテゴリー

被験者には、先述の各種素性の組合せによる SVM、タイトル類似度のみ、アソシエーションルールの確信度のみ、BOOKDB 要旨の類似度のみ、Amazon、のそれぞれによって推薦された図書の書誌事項を提示した。提示書誌事項は辻らと同様、図書のタイトル、著者、出版者、刊年の4つである。各図書が上記手法のいずれによって推薦されたかは被験者には分からない(そもそも推薦に上記手法を用いたことも被験者には知らせていない)。

被験者には、提示された書誌事項に基づいて、各図書の推薦の適切さを以下の5つのカテゴリーで判定してもらった。これも辻らと同様である：

2：「強い興味」

1：「読みたい」

0：「興味なし」

×：内容の判断ができないなどの理由で「評価不能」

△：既買って持っている、または読んだことがある

6. 結果

以下ではまず全体の結果を示した後、学年別、各手法の推薦値別、分野別、学習用データの多寡別の結果を示す。

6.1 全体の結果

先述の(1)各種素性の組合せによる SVM、(2)タイトル類似度のみ、(3)アソシエーションルールの確信度のみ、(4)BOOKDB 要旨の類似度のみ、(5)Amazon、のそれぞれによる推薦の評価結果は表1のようになった。上記(2)(3)(4)(5)による結果は、それぞれ表1の表側「タイトル」「貸出履歴」「BOOKDB」「Amazon」のもとに示してある。(1)については以下、煩雑さを避ける為、NDCの一致不一致、タイトル類似度、アソシエーションルールの確信度と支持度、BOOKDB 要旨の類似度、を素性とした SVM をそれぞれを「NDC」「タイトル」

November 2013

注ほか：図書館の貸出履歴と書誌情報を用いた図書推薦システムの有効性

表1：全体の結果

	△：既読	2：強い興味	1：読みたい	0,×：興味なし	合計
貸出履歴	5(3.9)	22(17.2)	41(32.0)	60(46.9)	128
タイトル	5(2.6)	34(17.7)	65(33.9)	88(45.8)	192
NDC+タイトル	3(3.1)	16(16.7)	41(42.7)	36(37.5)	96
タイトル+貸出履歴	3(3.1)	23(24.0)	35(36.5)	35(36.5)	96
NDC+タイトル+貸出履歴	6(6.3)	14(14.6)	49(51.0)	27(28.1)	96
BOOKDB+NDC+タイトル+貸出履歴	6(6.3)	12(12.5)	45(46.9)	33(34.4)	96
BOOKDB	0(0.0)	9(13.6)	32(48.5)	25(37.9)	66
Amazon	16(23.9)	22(32.8)	27(40.3)	2(3.0)	67

表2：大学院生に関する結果

	△：既読	2：強い興味	1：読みたい	0,×：興味なし	合計
貸出履歴	5(8.9)	6(10.7)	15(26.8)	30(53.6)	56
タイトル	4(5.1)	10(12.8)	25(32.1)	39(50.0)	78
NDC+タイトル	3(7.1)	4(9.5)	15(35.7)	20(47.6)	42
タイトル+貸出履歴	3(7.1)	7(16.7)	21(50.0)	11(26.2)	42
NDC+タイトル+貸出履歴	3(7.1)	4(9.5)	21(50.0)	14(33.3)	42
BOOKDB+NDC+タイトル+貸出履歴	3(7.1)	3(7.1)	20(47.6)	16(38.1)	42
BOOKDB	0(0.0)	2(8.3)	11(45.8)	11(45.8)	24
Amazon	13(34.2)	13(34.2)	11(28.9)	1(2.6)	38

表3：学部4年生に関する結果

	△：既読	2：強い興味	1：読みたい	0,×：興味なし	合計
貸出履歴	0(0.0)	7(15.2)	17(37.0)	22(47.8)	46
タイトル	0(0.0)	15(20.8)	28(38.9)	29(40.3)	72
NDC+タイトル	0(0.0)	7(19.4)	17(47.2)	12(33.3)	36
タイトル+貸出履歴	0(0.0)	11(30.6)	12(33.3)	13(36.1)	36
NDC+タイトル+貸出履歴	3(8.3)	5(13.9)	19(52.8)	9(25.0)	36
BOOKDB+NDC+タイトル+貸出履歴	3(8.3)	4(11.1)	17(47.2)	12(33.3)	36
BOOKDB	0(0.0)	4(13.3)	15(50.0)	11(36.7)	30
Amazon	2(10.5)	7(36.8)	10(52.6)	0(0.0)	19

表4：学部2年生に関する結果

	△：既読	2：強い興味	1：読みたい	0,×：興味なし	合計
貸出履歴	0(0.0)	9(34.6)	9(34.6)	8(30.8)	26
タイトル	1(2.4)	9(21.4)	12(28.6)	20(47.6)	42
NDC+タイトル	0(0.0)	5(27.8)	9(50.0)	4(22.2)	18
タイトル+貸出履歴	0(0.0)	5(27.8)	2(11.1)	11(61.1)	18
NDC+タイトル+貸出履歴	0(0.0)	5(27.8)	9(50.0)	4(22.2)	18
BOOKDB+NDC+タイトル+貸出履歴	0(0.0)	5(27.8)	8(44.4)	5(27.8)	18
BOOKDB	0(0.0)	3(25.0)	6(50.0)	3(25.0)	12
Amazon	1(10.0)	2(20.0)	6(60.0)	1(10.0)	10

「貸出履歴」「BOOKDB」を素性とした SVM と呼ぶことにし、素性の組合せについてはそれらを「+」で結んだ文字列で表すことにする。例えば「NDC+タイトル+貸出履歴」は NDC の一致不一致、タイトル類似度、アソシエーションルールの確信度と支持度、を素性とした SVM を表す。表 1 からこの「NDC+タイトル+貸出履歴」を用いて推薦した図書は96冊あり、そのうち「2：強い興味」と被験者に評価された図書は14冊で全体の14.6%(=14/96×100)を占めること、またタイトル類似度のみを用いて推薦した図書は192冊あり、そのうち「1：読みたい」と評価された図書は65冊で33.9%(=65/192×100)を占めること、などが分かる。⁵⁴⁾

さらと同様、「△：既読」「2：強い興味」「1：読みたい」の3つを「肯定的評価」とみなすと、表 1 から肯定的評価の割合は「NDC+タイトル+貸出履歴」の71.9%(=6.3+14.6+51.0)が最も高いことが分かる(ただし Amazon は除く)。次に高いのは「BOOKDB+NDC+タイトル+貸出履歴」の65.7%(=6.3+12.5+46.9)、次は「タイトル+貸出履歴」の63.6%(=3.1+24.0+36.5)である。これらの値は貸出履歴のみ(即ち、アソシエーションルールの確信度のみ)を用いた場合の53.1%(=3.9+17.2+32.0)、タイトル類似度のみを用いた場合の54.2%(=2.6+17.7+33.9)、BOOKDB 要旨の類似度のみを用いた場合の62.1%(=0.0+13.6+48.5)よりも高い。⁵⁵⁾従ってタイトルや貸出履歴といった情報を単独で用いるより、複数の情報を用いて SVM に学習させて推薦を行った方がパフォーマンスが高いことが言える。一方、タイトルのみや「NDC+タイトル」といった貸出履歴を用いない手法の肯定的評価は「NDC+タイトル+貸出履歴」より低く、「2：強い興味」は「タイトル+貸出履歴」より低い。従って利用者のプライバシー漏洩という危険性はあるものの、貸出履歴を図書推薦に援用することは、少なくともタイトルと NDC だけから推薦するより有効であると言える。

さて Amazon を除くと、「2：強い興味」という評価を受けた図書の割合は「タイトル+貸出履歴」の24.0%が最も高い。次に高いのは「タイトル」の17.7%、「貸出履歴」の17.2%、である。これらはいずれも NDC を参考にしていない点に注目したい。NDC を参考に推薦することは、今回の場合結果的に、現在興味ある1冊と書架上の位置に近い図書を

推薦することにつながりやすかった。この点については後述する。

最後に上記「2：強い興味」に関しても「肯定的評価」に関しても、本研究が提案する手法群はいずれも Amazon のそれに及ばなかった。Amazon の「2：強い興味」は32.8%であり、「肯定的評価」は97.0%である。この点は今後の課題としたい。

6.2 学年別の結果

被験者を大学院生、学部4年生、学部2年生に分けて評価を集計したところ表 2～表 4 のようになった。これらの表から、学部4年生と学部2年生においては、先述の肯定的評価の割合は「NDC+タイトル+貸出履歴」が最も高く、それぞれ75.0%(=8.3+13.9+52.8)、77.8%(=0.0+27.8+50.0)であることが分かる(ただし Amazon は除く)。⁵⁶⁾だが大学院生における「NDC+タイトル+貸出履歴」の肯定的評価の割合は66.6%(=7.1+9.5+50.0)で2位にとどまり、1位は「タイトル+貸出履歴」の73.8%(=7.1+16.7+50.0)となっている。また上記の肯定的評価の割合を並べると、大学院生(66.6%)<学部4年生(75.0%)<学部2年生(77.8%)、の順になっている。⁵⁷⁾即ち、学部生に関しては NDC を考慮した推薦の方が良いものの、大学院生に関しては NDC を考慮しない方が良い可能性がある。この点については次章で考察したい。またその為の補足データとして、これまで借りた冊数別に大学院生の「タイトル+貸出履歴」に対する肯定的評価の割合をまとめてみた。結果は表 5 のようになった。表の A～G は被験者を表している。表 5 から例えば大学院生被験者の中でこれまで借りた冊数が多い A は537冊を借りていること、借りた冊数が多い上位3名(=A～C)の肯定的評価は94.4%であることなどが分かる。全般に借りた冊数が多いグループの方が少ないグループよりも「タイトル+貸出履歴」

表 5：大学院生に関する「タイトル+貸出履歴」の結果

被験者	貸出冊数	肯定的評価(個別)	肯定的評価(二分)
A	537	100.0(=6/6)	94.4(=17/18)
B	108	83.3(=5/6)	
C	46	100.0(=6/6)	
D	28	100.0(=6/6)	58.3(=14/24)
E	19	50.0(=3/6)	
F, G	0	41.7(=5/12)	

November 2013

の肯定的評価が高い。上記94.4%に対して、借りた冊数が少ない4名(=D~G)の「タイトル+貸出履歴」の肯定的評価は58.3%にとどまっている。これらについて次章で考察する。⁵⁸⁾

さて「2:強い興味」の割合が最も高いのは、大学院生でも学部4年生でも先ほど同様「タイトル+貸出履歴」であった(それぞれ16.7%と30.6%)。学部2年生でも「2:強い興味」の割合は「タイトル+貸出履歴」が2位となっている(ただし同率2位が他に3つある)。

以上、「肯定的評価」に関しては、学部生では「NDC+タイトル+貸出履歴」が、大学院生では「タイトル+貸出履歴」が最も高いこと、「2:強い興味」に関しては、少なくとも大学院生と学部4年生では「タイトル+貸出履歴」が最も高いことが分かった。

6.3 推薦値別の結果

SVMを用いた場合は「所属確率」、タイトル類似度のみを用いた場合は「タイトル類似度」というように、本研究において推薦する図書の順位を決定した値をここでは推薦値と呼ぶことにする。6.1節で述べた全体結果は、推薦値が高い図書に評価を限定したら変わってくるかもしれない。そこでまず手法*i*が推薦した図書の集合を S_i 、 S_i の図書に付与された推薦値の集合を V_i 、 V_i の中央値を M_i とし、 S_i の図書をその推薦値が M_i 以上か M_i 未満かの2つに分割して、それぞれに関する評価結果をまとめてみた。結果は表6のようになった。表6から例えば「タイトル+貸出履歴」が推薦した図書を48冊ずつに分けた場合、推薦値が中央値以上の図書に関する肯定的評価は72.9%(=4.2+33.3+35.4)であるのに対し、中央値未満の図書に関する肯定的評価は54.2%(=2.1+14.6+37.5)にとどまることなどが分かる。中央値以上の図書群に関する評価結果を比較すると、肯定的評価の割合が最も高いのは「NDC+タイトル+貸出履歴」で75.0%(=10.4+18.8+45.8)であった(Amazonには推薦値に当たるものが表示されない為、表には含めていない)。また「2:強い興味」の割合が最も高いのは、これまで同様「タイトル+貸出履歴」で33.3%であった。⁵⁹⁾以上のことから、推薦値が高い図書に評価対象を限定しても「肯定的評価」は「NDC+タイトル+貸出履歴」が最も高く、「2:強い興味」は「タイトル+貸出履歴」

辻ほか:図書館の貸出履歴と書誌情報を用いた図書推薦システムの有効性が最も高いことが言えた。

6.4 分野別の結果

現在興味ある1冊のNDC(類網目の3桁)と、推薦図書のNDC(類網目の3桁)が同じ場合の評価結果と、異なる場合の評価結果を集計したところ、それぞれ表7,表8のようになった。表8から「NDC+タイトル+貸出履歴」はNDCが異なる図書は1冊も推薦しなかったことが分かる。即ち、「NDC+タイトル+貸出履歴」の高い肯定的評価は全てNDCが一致する図書によってもたらされたことが分かる。そして表7から、NDCが一致する図書を推薦した場合に最も肯定的評価が高いのは、これまで同様「NDC+タイトル+貸出履歴」であることが分かる(ただしAmazonは除く)。一方、「2:強い興味」の割合が表7,表8の中で最も高いのは、「タイトル+貸出履歴」がNDCの異なる図書を推薦した場合であることが分かる(27.7%。ただしAmazonは除く)。この点については次章で取り上げる。

現在興味ある1冊のNDCの類別に評価結果をまとめたところ表9~表12のようになった。表10,表12から3類と「0・3・4類以外」に関しては「NDC+タイトル+貸出履歴」の肯定的評価が最も高いことが分かる(ただしAmazonは除く)。また表11から4類に関しては「NDC+タイトル+貸出履歴」の肯定的評価は「タイトル+貸出履歴」に次いで2位であることが分かる。残る0類に関しては4位であるが、同率1位の3手法と有意な差はない。従って全般に、類別に見ても肯定的評価は「NDC+タイトル+貸出履歴」が最も高いように思われる。また「2:強い興味」の割合は、0類,3類,4類,「0・3・4類以外」の全てにおいて「タイトル+貸出履歴」が最も高いことが分かる(ただしAmazonは除く)。以上、現在興味ある1冊のNDCの類によって推薦手法を変えることの有効性は、今回の結果からは感じられず、肯定的評価と「2:強い興味」に関しては上記2手法を中心にするだけで十分であるように思われる。

6.5 学習用データの量別の結果

先述のように本研究では、186個の学習用データと358個の学習用データの2種類を用い、これまで後者の結果を示してきた。186個の学習用データを用いた結果は表13の「学習用データ小」のように

表6：各手法の推薦値が中央値以上の図書と中央値未満の図書に分けた結果

	中央値	△：既読	2：強い興味	1：読みたい	0,×：興味なし	合計
貸出履歴	以上	2(2.7)	15(20.3)	25(33.8)	32(43.2)	74
	未満	3(5.6)	7(13.0)	16(29.6)	28(51.9)	54
タイトル	以上	5(5.2)	16(16.7)	31(32.3)	44(45.8)	96
	未満	0(0.0)	18(18.8)	34(35.4)	44(45.8)	96
NDC+タイトル	以上	2(4.2)	9(18.8)	18(37.5)	19(39.6)	48
	未満	1(2.1)	7(14.6)	23(47.9)	17(35.4)	48
タイトル+貸出履歴	以上	2(4.2)	16(33.3)	17(35.4)	13(27.1)	48
	未満	1(2.1)	7(14.6)	18(37.5)	22(45.8)	48
NDC+タイトル+貸出履歴	以上	5(10.4)	9(18.8)	22(45.8)	12(25.0)	48
	未満	1(2.1)	5(10.4)	27(56.3)	15(31.3)	48
BOOKDB+NDC+タイトル+貸出履歴	以上	1(2.1)	6(12.5)	21(43.8)	20(41.7)	48
	未満	5(10.4)	6(12.5)	24(50.0)	13(27.1)	48
BOOKDB	以上	0(0.0)	5(15.2)	19(57.6)	9(27.3)	33
	未満	0(0.0)	4(12.1)	13(39.4)	16(48.5)	33

表7：現在興味ある1冊のNDCと推薦図書のNDCが同じだった場合の結果

	△：既読	2：強い興味	1：読みたい	0,×：興味なし	合計
貸出履歴	2(5.4)	8(21.6)	16(43.2)	11(29.7)	37
タイトル	2(3.2)	14(22.2)	23(36.5)	24(38.1)	63
NDC+タイトル	3(3.6)	14(16.7)	38(45.2)	29(34.5)	84
タイトル+貸出履歴	1(2.2)	10(21.7)	22(47.8)	13(28.3)	46
NDC+タイトル+貸出履歴	4(4.7)	13(15.3)	44(51.8)	24(28.2)	85
BOOKDB+NDC+タイトル+貸出履歴	3(4.1)	10(13.7)	39(53.4)	21(28.8)	73
BOOKDB	0(0.0)	2(10.5)	9(47.4)	8(42.1)	19
Amazon	6(17.6)	9(26.5)	18(52.9)	1(2.9)	34

表8：現在興味ある1冊のNDCと推薦図書のNDCが異なった場合の結果

	△：既読	2：強い興味	1：読みたい	0,×：興味なし	合計
貸出履歴	3(3.3)	14(15.4)	25(27.5)	49(53.8)	91
タイトル	3(2.5)	19(15.6)	38(31.1)	62(50.8)	122
NDC+タイトル	0(0.0)	0(0.0)	1(50.0)	1(50.0)	2
タイトル+貸出履歴	1(2.1)	13(27.7)	13(27.7)	20(42.6)	47
NDC+タイトル+貸出履歴	0(-)	0(-)	0(-)	0(-)	0
BOOKDB+NDC+タイトル+貸出履歴	1(6.7)	1(6.7)	5(33.3)	8(53.3)	15
BOOKDB	0(0.0)	6(14.0)	20(46.5)	17(39.5)	43
Amazon	10(31.3)	12(37.5)	9(28.1)	1(3.1)	32

なった。この表から学習用データが少なかった場合、「NDC+タイトル+貸出履歴」の肯定的評価は54.1% (=2.6+18.2+33.3)にとどまり、先述の71.9%よりかなり低いことが分かる。「2：強い興味」の割合は、「NDC+タイトル+貸出履歴」「BOOKDB+

NDC+タイトル+貸出履歴」においては学習用データが少ない方が高いものの、「タイトル+貸出履歴」においては学習用データが多い方が高い(24.0%と12.0%)⁶⁰⁾。学習用データを約2倍にただけで上記の差が見られたことから、今後被験者を集めてより

November 2013

辻ほか：図書館の貸出履歴と書誌情報を用いた図書推薦システムの有効性

表9：現在興味ある1冊が0類だった場合の結果

	△：既読	2：強い興味	1：読みたい	0,×：興味なし	合計
貸出履歴	0(0.0)	3(9.1)	14(42.4)	16(48.5)	33
タイトル	4(8.3)	6(12.5)	20(41.7)	18(37.5)	48
NDC+タイトル	3(10.0)	5(16.7)	13(43.3)	9(30.0)	30
タイトル+貸出履歴	2(6.7)	6(20.0)	14(46.7)	8(26.7)	30
NDC+タイトル+貸出履歴	3(10.0)	3(10.0)	15(50.0)	9(30.0)	30
BOOKDB+NDC+タイトル+貸出履歴	1(3.3)	4(13.3)	17(56.7)	8(26.7)	30
BOOKDB	0(0.0)	5(16.7)	17(56.7)	8(26.7)	30
Amazon	5(22.7)	6(27.3)	11(50.0)	0(0.0)	22

表10：現在興味ある1冊が3類だった場合の結果

	△：既読	2：強い興味	1：読みたい	0,×：興味なし	合計
貸出履歴	0(0.0)	8(22.2)	13(36.1)	15(41.7)	36
タイトル	1(1.9)	10(18.5)	20(37.0)	23(42.6)	54
NDC+タイトル	0(0.0)	4(13.3)	13(43.3)	13(43.3)	30
タイトル+貸出履歴	0(0.0)	8(26.7)	10(33.3)	12(40.0)	30
NDC+タイトル+貸出履歴	0(0.0)	5(16.7)	15(50.0)	10(33.3)	30
BOOKDB+NDC+タイトル+貸出履歴	0(0.0)	3(10.0)	16(53.3)	11(36.7)	30
BOOKDB	0(0.0)	1(8.3)	5(41.7)	6(50.0)	12
Amazon	1(6.3)	8(50.0)	6(37.5)	1(6.3)	16

表11：現在興味ある1冊が4類だった場合の結果

	△：既読	2：強い興味	1：読みたい	0,×：興味なし	合計
貸出履歴	5(17.2)	5(17.2)	4(13.8)	15(51.7)	29
タイトル	0(0.0)	8(22.2)	10(27.8)	18(50.0)	36
NDC+タイトル	0(0.0)	2(16.7)	4(33.3)	6(50.0)	12
タイトル+貸出履歴	1(8.3)	4(33.3)	7(58.3)	0(0.0)	12
NDC+タイトル+貸出履歴	0(0.0)	2(16.7)	9(75.0)	1(8.3)	12
BOOKDB+NDC+タイトル+貸出履歴	2(16.7)	2(16.7)	5(41.7)	3(25.0)	12
BOOKDB	0(0.0)	1(16.7)	2(33.3)	3(50.0)	6
Amazon	9(52.9)	3(16.7)	4(23.5)	1(5.9)	17

表12：現在興味ある1冊が0・3・4類以外だった場合の結果

	△：既読	2：強い興味	1：読みたい	0,×：興味なし	合計
貸出履歴	0(0.0)	6(20.0)	10(23.3)	14(46.7)	30
タイトル	0(0.0)	10(18.5)	15(27.8)	29(53.7)	54
NDC+タイトル	0(0.0)	5(20.8)	11(45.8)	8(33.3)	24
タイトル+貸出履歴	0(0.0)	5(20.8)	4(16.7)	15(62.5)	24
NDC+タイトル+貸出履歴	3(12.5)	4(16.7)	10(41.7)	7(29.2)	24
BOOKDB+NDC+タイトル+貸出履歴	3(12.5)	3(12.5)	7(29.2)	11(45.8)	24
BOOKDB	0(0.0)	2(11.1)	8(44.4)	8(44.4)	18
Amazon	1(8.3)	5(41.7)	6(50.0)	0(0.0)	12

表13：学習用データの量による違い

		△：既読	2：強い興味	1：読みたい	0, X：興味なし	合計
学習用 データ小	タイトル+貸出履歴	0(0.0)	23(12.0)	56(29.2)	113(58.9)	192
	NDC+タイトル+貸出履歴	5(2.6)	35(18.2)	64(33.3)	88(45.8)	192
	BOOKDB+NDC+タイトル+貸出履歴	4(2.1)	37(19.3)	66(34.4)	85(44.3)	192
学習用 データ大 (再掲)	タイトル+貸出履歴	3(3.1)	23(24.0)	35(36.5)	35(36.5)	96
	NDC+タイトル+貸出履歴	6(6.3)	14(14.6)	49(51.0)	27(28.1)	96
	BOOKDB+NDC+タイトル+貸出履歴	6(6.3)	12(12.5)	45(46.9)	33(34.4)	96

規模の大きい学習用データを作成し使用することで、「NDC+タイトル+貸出履歴」の肯定的評価や「タイトル+貸出履歴」の「2：強い興味」の割合は、本研究の値よりも高く出来る可能性がある。

7. 考察

本研究では、肯定的評価の割合は「NDC+タイトル+貸出履歴」が最も高く、「2：強い興味」の割合は「タイトル+貸出履歴」が最も高いという結果が示された。6.4節で述べたように「NDC+タイトル+貸出履歴」はNDCが「現在興味ある1冊」と同じ図書しか推薦せず、「タイトル+貸出履歴」は異なる図書も推薦していた。以下ではこの点に焦点を当てた検討を行いたい。

図書館の書架は一種の図書推薦の場と言える。研究テーマや関心が固まっており図書館でよく図書を借りる利用者、いわば行きつけの書架がある利用者は、書架による推薦を通じて自分の興味と関わる図書を、ある程度把握出来ている可能性がある。従って、現在興味ある1冊とNDCが同じ図書(即ち、行きつけの書架にある図書)には、既に知っている図書が多かったのかもしれない⁶¹⁾。6.2節で示したように、(1)「NDC+タイトル+貸出履歴」に関する肯定的評価の割合が、大学院生<学部4年生<学部2年生、の順になっていたこと、(2)大学院生に関してはNDCが異なる図書も推薦する「タイトル+貸出履歴」の方が、NDCが同じ図書しか結果的に推薦しなかった「NDC+タイトル+貸出履歴」より肯定的評価の割合が高かったこと、(3)「タイトル+貸出履歴」の肯定的評価の割合は、これまで借りた冊数が多いグループの方が、少ないグループよりも高かったこと、は上記推定の正しさを裏付けているように思われる。このように考えると、学年やこれまで借りた冊数などから行きつけの書架を持つかを推定し、持たない利用者に対しては「NDC+

タイトル+貸出履歴」で推薦を行い、持っていないような利用者には「タイトル+貸出履歴」で推薦を行うというように、利用者によって両手法を切り替えることも考えられる。

次に本研究ではBOOKDBの要旨の類似度も用いたが、期待したほどの効果は見られなかった。図書館の蔵書には今回用いたBOOKDB2005年版～2011年版に含まれないものが多かったことがパフォーマンス向上に寄与しなかった理由の1つかもしれない。過去のバージョンを買い進め、網羅性を高めればあるいは有効かもしれないが、BOOKDBの価格は1年分が約40万円と高価である為⁶²⁾、実施には難しい面が多い⁶³⁾。

さて辻らと同様、今回もAmazonのパフォーマンスを上回ることは出来なかった。Amazonを超えるには、6.5節で述べたように学習用データを増やしながら、かつ大学図書館や大学生といった想定環境や想定利用者に特化した推薦を実現していくことが有効と思われる。例えば、学生が学籍番号を入力するとその学生の履修科目群が情報としてシステムに入力され、シラバスに挙げられている各科目の参考書・教科書等をその旨表示しながら推薦する、などである。もっともAmazonを超えることに固執するのではなく、辻らも述べているように、Amazonが推薦する図書をそのままOPACに表示するという低コストな推薦方法も検討する必要がある。またAmazonの推薦情報そのものをSVMの素性の1つに組み込んでしまうことも考えられる。

最後に誤り分析の結果を述べる。(1)本研究では単純なコサイン尺度をタイトル類似度としたが、この尺度は長いタイトルよりも短いタイトルの方が「似ている」という結果を出しやすい。そして最近の図書のタイトルには長いものが多く、逆に昔の図書にはシンプルで短いものが多いことから、タイトル類似度による推薦はかなり昔の図書を推薦する傾

November 2013

向があった。被験者からも「もっと新しい図書を推薦してほしい」という意見があったので、今後は刊行年も素性として SVM に組み込むことが考えられる。(2) また MeCab が漢字数字や○を1単語として切ってしまうことに由来する誤りも見られた。例えば「一〇〇年」を「一」「〇」「〇」「年」という4単語に切ってしまう為、タイトル中の他の単語の違いがあまり反映されず、「一〇〇年」という語を共通に含む図書だけを推薦するといった誤りがあった。これらは MeCab の辞書を修正することで対応出来ると思われる。

8. おわりに

これまで貸出履歴を用いた協調フィルタリングやアソシエーションルールによって図書推薦を行う研究はあったが、タイトルや NDC など図書推薦に有効と思われる貸出履歴以外の情報を併用する研究はほとんど行われてこなかった。それに対し本研究では、そうした複数の情報を SVM によって統合的に利用する方法及びその有効性を示した。今回は貸出履歴、タイトル、NDC、BOOKDB の要旨を用いたが、本手法はこれら以外の情報も自由に組み込むことが出来る点で、今後の拡張性・発展性が高いものである。

本研究では「NDC+タイトル+貸出履歴」や「タイトル+貸出履歴」といった複数の情報を SVM で統合的に用いた手法が、貸出履歴単独やタイトル類似度単独よりも、あるいは他の情報の組合せよりも、高いパフォーマンスで図書を推薦することを示した。従って図書推薦に貸出履歴を用いることは、少なくともタイトルと NDC だけで図書推薦を行うより有効であると言える。貸出履歴の利用はプライバシー漏洩の危険を伴うとはいえ、今後参考にすべき結果と言えよう。また本研究では、学年が上がるにつれ、NDC が異なる図書を推薦した方が相対的に高い評価が得られる可能性を示した。対象利用者によって推薦手法を切り替えることも検討する必要がある。

今後はまず、学習用データを増やして行った場合に、本研究手法のパフォーマンスがどこまで向上するかを検証したい。また前章で述べたように各学生の履修科目とその参考書といった大学生特有の情報を図書推薦に組み込む可能性を追求したい。さらに本研究ではいわば「興味の度合い」しか被験者に聞いておらず、被験者がどのように感じて興味がある

辻はか：図書館の貸出履歴と書誌情報を用いた図書推薦システムの有効性と回答したのかを考慮していない。例えば思いがけない新分野の図書を推薦されて興味があると回答したのか、既知文献だが未読なので興味があると答えたのかなどである。今後は興味の類型化を進め、それぞれに最適な図書推薦手法の開発を進めたい。最後に本研究では自動分類の代表的手法である SVM のみを取り上げたが、自動分類には他にランダムフォレストやナイーブベイズといった手法がある。それらが SVM のパフォーマンスを著しく上回る可能性は低いと思われるが、今後比較検証を進めたい。

謝辞

実験に協力して下さった被験者の皆様、貸出履歴データを提供して下さった T 大学図書館の皆様に、この場を借りて御礼申し上げます。本当にありがとうございました。

注・引用文献

- 1) 安東奈穂子ら「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」『デジタル図書館』(30), 2006, p.62-71.
- 2) 岡本真「Web2.0時代の図書館-Blog, RSS, SNS, CGM」『情報の科学と技術』56(11), 2006, p.502-508.
- 3) 渡邊齊志「知的自由の陥穽：利用情報保護思想が公立図書館に及ぼす影響の分析」『Library and Information Science』(58), 2007, p.103-115.
- 4) 佐浦敬之, 辻慶太「公共図書館における利用履歴の活用に関する意識調査」『第57回日本図書館情報学会研究大会発表要綱』, 2009, p.1-4.
- 5) 小野永貴, 常川真央「Web 時代にあるべき未来の図書館サービスの胎動：貸出履歴の議論を超えた Shizuku2.0の実現へ」『情報管理』53(4), 2010, p.185-194.
- 6) 原田隆史「図書館の貸出履歴を用いた図書の推薦システム」『デジタル図書館』(36), 2009, p.22-31.
- 7) 原田隆史, 増田浩佑「貸出記録を用いた図書推薦システムにおける重み付けの変更」『デジタル図書館』(38), 2010, p.54-66.
- 8) Keita Tsuji et al., "Use of Library Loan Records for Book Recommendation," *Proceedings of the 3rd IIAI International Conference on e-Services and Knowledge Management (IIAI ESKM 2012)*, 2012, p.30-35.
- 9) 辻慶太ら「図書館の貸出履歴を用いた図書推薦システムの有効性検証」『図書館界』, 64(3), 2012, p.176-189.
- 10) 「BOOK」データベース。
〈<http://www.nichigai.co.jp/dcs/index3.html>〉。[引用日：2013-05-10]
- 11) Amazon. 〈<http://www.amazon.co.jp/>〉。[引用日：2013-

- 05-10]
- 12) 斉藤国博, 中條将典「おすすめ商品を見つける仕組み: 「Amazon.co.jp」(アマゾンジャパン) など: どうなっているの? あのソフトの仕組み」『日経 ITPro』2006.
 <<http://itpro.nikkeibp.co.jp/article/COLUMN/20061012/250605/?ST=develop&P=3>>. [引用日: 2013-05-10]
- 13) Amazon Product Advertising API.
 <<https://affiliate.amazon.co.jp/gp/advertising/api/detail/main.html>>. [引用日: 2013-05-10]
- 14) 原田隆史, 前掲 6).
- 15) 原田隆史, 増田浩佑, 前掲 7).
- 16) Tsuji et al., *op. cit.* 8).
- 17) 辻慶太ら, 前掲 9).
- 18) Colleen Whitney and Lisa Schiff “Melvil Recommender Project: Developing Library Recommendation Services,” *D-Lib Magazine*, 12(2), 2006.
 <<http://www.dlib.org/dlib/december06/whitney/12whitney.html>>. [引用日: 2013-05-10]
- 19) Saurabh Shirgaonkar et al., “Application of Improved Apriori in University Library,” *International Conference and Workshop on Emerging Trends in Technology (ICWET2010)*, 2010, p.535-540.
- 20) Chia-Chen Chen and An-Pin Chen “Using Data Mining Technology to Provide a Recommendation Service in the Digital Library,” *The Electronic Library*, 25(6), 2007, p.711-724.
- 21) Yongcheng Luo et al., “A Privacy-Preserving Book Recommendation Model Based on Multi-Agent,” *Proceedings of the 2009 Second International Workshop on Computer Science and Engineering*, 2009, p.323-327.
- 22) 神島敏弘「推薦システムのアルゴリズム(1)」『人工知能学会誌』22(6), 2007, p. 826-837.
- 23) 神島敏弘「推薦システムのアルゴリズム(2)」『人工知能学会誌』23(1), 2008, p. 89-103.
- 24) 神島敏弘「推薦システムのアルゴリズム(3)」『人工知能学会誌』23(2), 2008, p. 248-263.
- 25) Dietmar Jannach ら(田中克己, 角谷和俊監訳)『情報推薦システム入門: 理論と実践』共立出版, 2012, 359p.
- 26) T 大学は日本有数の研究系総合大学である。本研究で得られた傾向は他のタイプの大学では観察されないかもしれない。この点の検証は今後の課題としたい。
- 27) その他の利用者は専攻分野など各種属性が把握しづらいため除外することにした。
- 28) 「BOOK」データベース, 前掲 10).
- 29) こちらの貸出履歴には, 前掲 9) の辻らのものを用いた。
- 30) こちらの貸出履歴には今回のもの, 即ち先述の2006年1月2日から2012年3月31日までのものを用いた。
- 31) ただし一度も借りられていない図書などもあるので, アソシエーションルールが各被験者に推薦する図書は6冊に満たないことがあった。また Amazon が「その図書を購入した人は以下の図書も購入しています」と提示する図書が T 大学図書館にないこともあるので, Amazon についても各被験者に推薦する図書が6冊に満たないことがあった。
- 32) 前掲 9) の辻らの被験者33名と本研究の被験者32名のうち大学院生5名は重複している(学部2年生と4年生に重複はない)。従って前者の評価結果を学習用データに用い, 重複する学生に評価を求めたのは問題があったかもしれない。ただ, グループ T に含まれ, 評価まで求めたのは5名中3名であり, また前回の実験から1年が経過しており, この3名の現在興味ある1冊も以下のように異なっていた。いずれも NDC の類のレベルで異なっており, 推薦した図書にも重なりはなかった。以上のことから影響は小さいと思われる。
- 以下は上記3名に関する前掲 9) での現在興味ある1冊のタイトル, 刊年, 出版者, NDC→本研究での現在興味ある1冊のタイトル, 刊年, 出版者, NDC, である:
- ・『本の現場: 本はどう生まれ, だれに読まれているか』, 2009, ポット出版, 020→『リサーチ・マインド経営学研究法』, 2005, 有斐閣, 335.
 - ・『Ruby 逆引きハンドブック』, 2009, シーアンドアール研究所, 007→『これなら分かる最適化数学: 基礎原理から計算手法まで』, 2005, 共立出版, 417.
 - ・『公共部門の業績評価: 官と民の役割分担を考える』, 2005, 東京大学出版会, 343→『新たな図書館・図書館史研究: 批判的図書館史研究を中心にして』, 2011, 日本図書館協会, 010.
- 33) MeCab: Yet Another Part-of-Speech and Morphological Analyzer. Software available at
 <<http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html>>. [引用日: 2012-12-12]
- 34) 単名詞, 2名詞列とはそれぞれ1つの名詞, 連続する2つの名詞のことである。例えば「特許」「情報」「検索」という3つの名詞が連続する部分からは「特許」「情報」「検索」という3つの単名詞と「特許情報」「情報検索」という2つの2名詞列を抽出することになる。
- 35) BOOKDB の類似度だけで推薦する場合, 推薦する図書の集合は BOOKDB に要旨が挙げられているものに実質上限定される。だが推薦図書を選ぶ集合が手法によって変わることは往々にしてあるので本研究では特に問題視しない。
- 36) 辻慶太ら, 前掲 9)。
- 37) 福田剛志ら『データマイニング』共立出版, 2001, p.19-32. (データサイエンス・シリーズ3)。
- 38) 金明哲「フリーソフトによるデータ解析・マイニング第40回: アソシエーション分析(1)」『ESTRELA』(152), 2006, p. 52-57.
- 39) その確かさについては後述する。また「図書 B, C を同時に借りる利用者は図書 A も借りる」というルールを引き出すことも出来るが, 本研究ではこのように3冊以上を同

November 2013

時に扱うルールは取り上げない。

- 40) 確信度と支持度に関するこれらの閾値は前掲 9) と同じである。支持度が0.000001以上というのは本研究のデータにおいては1回以上貸し出されたことを意味する。前掲 9) では他にリフトも考慮したが、本研究では学習用データの少なさを考慮して採用しなかった。従って本研究における貸出履歴に基づく推薦は、前掲 9) より若干シンプルな形になっている。この点は今後の課題としたい。だが、後述の肯定的評価を前掲 9) のそれと比較すると、結果に大きな差はないことが分かる。
- 41) 貸出履歴に一度も現れない図書の支持度は当然0であるが、確信度は算出することができない。このような場合本研究では確信度も便宜上0とした。この点は今後の課題としたい。
- 42) これらの組合せを選ぶに当たっては、評価する被験者の負担をかなり意識した。本来は全ての組合せを試すべきだったかもしれない。だが確信度と支持度だけでは推薦図書が6冊に満たない場合が多く、それらとNDCの一致不一致だけを組み合わせても良い結果は生まないと思われる。そしてBOOKDBも「現在興味ある1冊」を含まないことが多く、その場合BOOKDBは推薦図書を増やす方向には機能しない為、確信度・支持度とBOOKDBだけを組み合わせても良い結果は生まないと思われる。
- 43) 類1桁だけあるいは類網2桁なども試したが、3桁が最も有効そうであった為3桁を採用した。
- 44) Chih-Chung Chang and Chih-Jen Lin “LIBSVM: A Library for Support Vector Machines,” *ACM Transactions on Intelligent Systems and Technology*, 2(3), 2011, p.27:1-27:27. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>. [引用日: 2012-11-03]
- 45) 阿部重夫『パターン認識のためのサポートベクトルマシン入門』森北出版株式会社, 2011, p.22-31.
- 46) このChang and Lin, *op. cit.* 44). の式は阿部 *ibid.*, p. 23. の(2.38)式と異なっているが、これはChang and Linがカーネル関数を構成する $\phi(x_i)$ をあらかじめ式内に組み込んでいる為であり、最終的には両者は等価である。
- 47) `easy.py` はLIBSVMに標準で添付されている。
- 48) Chang and Lin, *op. cit.* 44).
- 49) 阿部重夫, 前掲 45).
- 50) Ting-Fan Wu, Chih-Jen Lin and Ruby C. Weng “Probability Estimates for Multi-class Classification by Pairwise Coupling,” *Journal of Machine Learning Research*, 5, 2004, p.975-1005.
<<http://www.csie.ntu.edu.tw/~cjlin/papers/svmprob/svmprob.pdf>>. [引用日: 2013-05-10]
- 51) Amazon が6冊提示しなかった場合は、提示した5冊以下の図書のうちT大学図書館が所蔵している図書のみを推薦した。Amazonは利用者のディスプレイによっては7冊

はか: 図書館の貸出履歴と書誌情報を用いた図書推薦システムの有効性

以上提示することもあるが、6冊までしか提示しないディスプレイを本研究では使用した。

- 52) Amazon Product Advertising API, 前掲 13).
- 53) 前掲 6) の原田は、被験者の専攻分野を人文科学, 社会科学, 自然科学, 図書館情報学の4つに分け、分野間で評価結果に有意差は見られなかったとしている。そこで本研究でも被験者の専攻分野は結果に大きな影響を与えないと仮定する。だが原田が用いたサンプルは少ない為、この点は今後の課題としたい。
- 54) 被験者により「×: 評価不能」とされた図書は、タイトルや著者名だけでは少なくとも被験者の興味を惹かなかつた図書であることから、「0: 興味なし」と合わせて集計してある。
- 55) 「NDC+タイトル+貸出履歴」の71.9%と「BOOKDB+NDC+タイトル+貸出履歴」の65.7%の間に有意差は認められないが、前者の71.9%と貸出履歴のみの53.1%, タイトル類似度のみの54.2%との間には有意水準0.05で差が認められた。
- 56) ただし学部2年生に関しては「NDC+タイトル」も同率1位である。
- 57) ただしサンプルが少ないこともあり、統計的な有意差はない。
- 58) 学部4年生, 2年生で最も図書を借りている人の冊数はそれぞれ53冊, 35冊と少ないからか、学部生に関しては借りた冊数の多寡で被験者を2分しても上記のような結果にはならなかった。
- 59) ただし有意水準0.05では2位との間に有意差は認められない。
- 60) これら71.9%と54.1%, 24.0%と12.0%の間にはそれぞれ有意水準0.05で差が認められた。
- 61) 実際「NDC+タイトル+貸出履歴」の方が「タイトル+貸出履歴」より「△: 既読」の割合が高い。
- 62) 「BOOK」データベース, 前掲 10).
- 63) 本研究では図書の類似度の計算にのみBOOKDBの要旨を用いたが、要旨については利用者への推薦時に参考情報として提示するなど様々な使い方が考えられる。これについては今後の課題としたい。