# 平成 29 年度
# 国立情報学研究所教育研修事業
# 学術情報システム総合ワークショップ

## 成果報告書

## 「相対的入門書判別器の開発」

国立国会図書館 電子情報部電子情報サービス課次世代システム開発研究室　青池 亨
国際日本文化研究センター 情報管理施設 資料課 資料利用係　荒木 のりこ
東京大学 情報システム部情報基盤課学術情報チーム　石田 唯
情報・システム研究機構 国立情報学研究所 学術基盤推進部学術コンテンツ課　瀬尾 崇一郎
早稲田大学 図書館総務課　長谷川 敦史

# 目次

# 1. はじめに

平成 29 年度の学術情報システム総合ワークショップのテーマは、「AI 技術の理解とサービス・業務への適用」である。テーマに基づき、現在の業務上の課題について検討した結果、選書・レファレンス・分類業務に役立つシステム「相対的入門書判別器」を企画・開発した。

## 1.1. 背景

大学図書館の書籍は専門性の強い書籍から初心者向けの入門書まで、その難易度は多様である。特にある分野の初心者にとって、入門書には特別な需要が存在する。また、図書館員にとっても、多分野にわたる大量の書籍の中から入門書を判別しレファレンスすることは困難を伴う。

現在、図書館員は書誌情報・レビュー・シラバス等の情報と自身の経験に基づいて選書を行っているが、AI 技術を用いればそれと同等の結果を得ることができるのではないだろうか。また、システム化することでユーザが使いやすいサービスを提供することができるのではないだろうか。

## 1.2. 目的・課題

ユーザが内容に興味があって手に取り、「読んでみたけど読めなかった」書籍があるとする。それは、①表現の難易度が高かった（表現の難易度）、②専門的な内容についての基礎知識が足りなかった（内容の基礎的さ）といったことに原因がある。

そこで、ユーザに「読めなかった」1 冊の ISBN もしくはキーワードを入力してもらい、それと同様の内容を持つ資料群をスコアリングして、①や②の問題を解決できる資料を提示するシステムを企画・開発することを目的とした。

上記のシステムを開発するため、下記の点について検討した。

- 入門書とはどのようなものかを定義づける。
- ユーザが入力する 1 冊と同等の内容を持つ資料群を出力する。
- その資料群の中でスコアリングする。
- スコアリングを元にユーザが入力したものよりもより入門書といえる書籍を提示する。

## 2. 議論の過程

### 2.1. 第 1 回集合研修（7/6-7/7）

#### 2.1.1. 1 日目テーマ決定～テーマ発表

AI 技術に関する講義・事例紹介の後、それをどのように図書館システムに取り入れることができるか、現在の課題を挙げながら話し合った。さまざまな意見があった中で、ユーザからの需要が高いものの、レファレンス時の対応が難しい「入門書の推薦」をテーマとすることに決定した。

#### 2.1.2. 2 日目討議～発表

入門書を推薦するにあたって、入門書かどうかを人はどうやって見分けているのか、機械的に処理ができるのかが問題となった。

- ある書誌について、それが入門書であるかどうか、どの程度入門書らしいかを判定する。
- 「明確に入門書である」と見なせるような書誌データを集め、これを教師データとした機械学習の可能性を検討する。
- 人間はどのようにして入門書を見つけるのか。

（質問・意見）

- 「入門書」とは？「入門書ではない」とは？
  ー「入門書ではないもの」からのアプローチによる判別も必要なのではないか。
- なぜ判別器を作るのか、既存の事例との違いはあるのか。
  回答：レファレンス協同データベースやシラバスのデータを使った例は珍しいと思われる。権威づけを行えるので信頼性の高いデータになるのではないか。また、図書館員が自館の蔵書構成を把握し切れていない中で、需要のある入門書をターゲットにした。
- 入門書を置いてあるフロアで判別できるかもしれない。また、入門書は複本率が高くなりやすい。
- 図書館員がユーザに提供するリストを作るときはテスト前の貸出冊数や予約冊数も見ている。
- 丸善の選書カタログに難易度を示す★がついている。
- 最終的なサービスの形はどんなものになるか。
  回答：上から順番に適切なものが並んでいるもの。ジャンルを入れると返すもの。または ISBN を入れると初心者向けかどうか返すもの。

### 2.2. 集合研修後（7/8-8/30）

状況は都度 backlog 上で共有し、コメントを連ねる形で相談した。課題は以下の通りで

ある。

・書誌情報のかたまりを取得・整形（NDL サーチ API）
・ツールの調査（使用方法、どのような情報を手に入れることができるか）
－商業 API（Amazon・Yahoo!・楽天・Google books・openBD）
－Wikipedia API・Google Custom Search API
－図書館関係の API（CiNii Books・レファレンス協同データベース・カーリル）
・先行研究・事例の調査

## 2.3. 第 2 回集合研修（8/31-9/1）

### 2.3.1. 1 日目発表

前回研修以降に得られた情報・議論した内容をまとめた。また、第 1 回 2 日目のグループ発表時に提示された、「入門書とは何か」という根本的な問題についてさらに討議を進めた。

・入門書とは何か。
・API でどのような情報を得ることができるか。
・実験：複本率・件数
－十進分類上の分類を条件に取得した書籍を対象に、複本率による判定を検証
－「社会学」と「機械工学」を対象とした。
－NDL サーチでそれぞれの書籍リストを取得し、CiNii Books で所蔵館を調べ、それぞれの複本状況（同じ機関で複数の図書館・室が所蔵）を調査し、ソートした。
・シラバスに掲載されている書籍を取得するには、どうしたらよいか？
－宇都宮大学附属図書館などは OPAC 上で検索可能である。

（質問・意見）

・情報として今取れないから使わない、というのは理由として良くない。本当に使いたいデータなら、どうやって得られるか考えるべきである（レビュー・全文など）。
・アウトプットについて
－「社会学の入門書はこれら」と群で示すのか。
－「この本は入門書？」に対して yes no で答えるのか。
－「この本より簡単な本は？」に対して相対的な判定を出すのか。
→ユーザからすると相対的なものがありがたい。

### 2.3.2. 1 日目討議～2 日目発表

システム開発にあたり、具体的にどういった手順・作業が必要か討議した。また、シス

テムの動きを図式化した。

・システムの動き

－ユーザが提示した書籍と同じ内容のものを抽出する。

－書籍ごとの「入門書らしさ」をスコアリングする。

－ユーザが提示した書籍よりも「入門書らしさ」が高いものを推薦する。

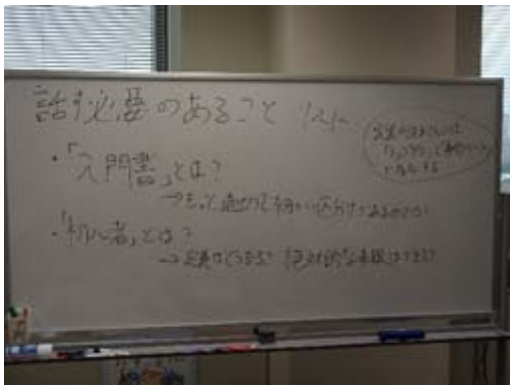



図 2.3-1 話す必要のある事

・入門書とは何か。

・初心者（入門書を求める人）とは何か。

－本システムにおける、「入門書」と「初心者」の定義について話し合った。まず「初心者」は、ある本を読もうとしたが難しくて読めなかった人、とした。本システムを利用するユーザである。そして、「入門書」は、「初心者」が読むことで、難しくて読めなかった本を読めるようになるような本、とした。

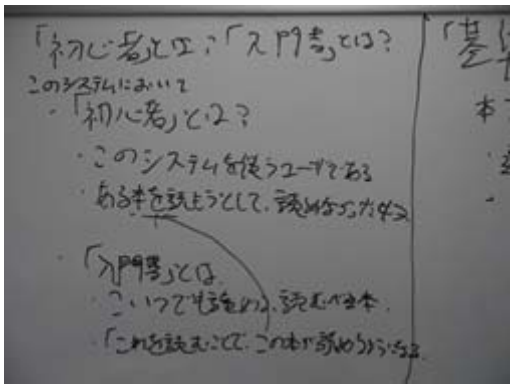



図 2.3-2 「初心者」とは？「入門書」とは？

・最終的にどのようにユーザに提供するのか。

－入門書のリストを提示するシステム

－入門書かどうかを判定する API

－サンプル本からの相対的入門書を提示するシステム

議論の結果、「サンプル本からの相対的入門書を提示するシステム」を作成することとした。

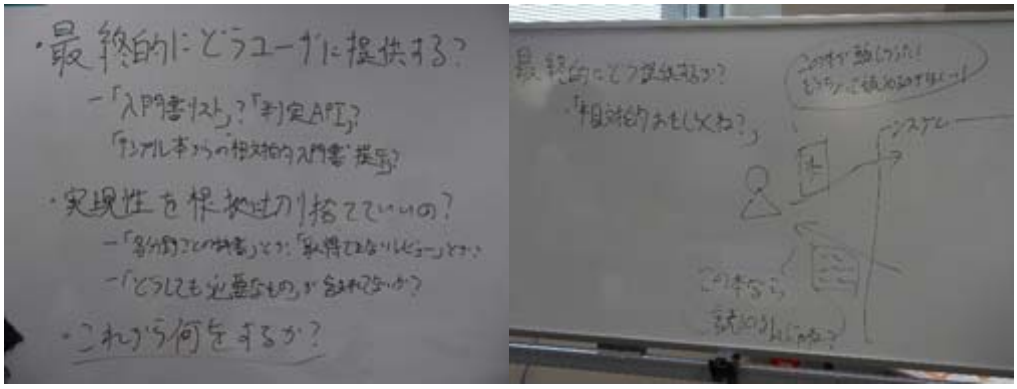



図 2.3-3 ユーザへの提供形式

・ユーザが提示した書籍と同じ内容のものの抽出

ユーザが提示した書籍（「基準本」）と同じ内容のものを抽出する方法として、連想検索、目録（分類）、件名（上位下位を含む）といった案があがった。

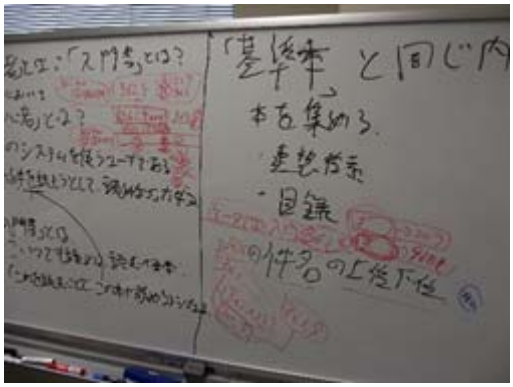



図 2.3-4 「基準本」と同じ内容の本を抽出

・書籍ごとの「入門書らしさ」のスコアリング

まず、「入門書らしさ」を判定する基準について話し合った。その結果、「難しくて読めなかった」原因は 2 種類に分けられるという結論に至った。「表現の難易度」と「内容の基礎的さ」である。「表現の難易度」については、タイトル、目次、帯、内容紹介などのテキスト的難易度から判定できるのではないかと考えた。また、表紙画像からも難易度を推定できないか、という案があがった。「内容の基礎的さ」については、所蔵情報から導き出すことができるのではないかと考えた。

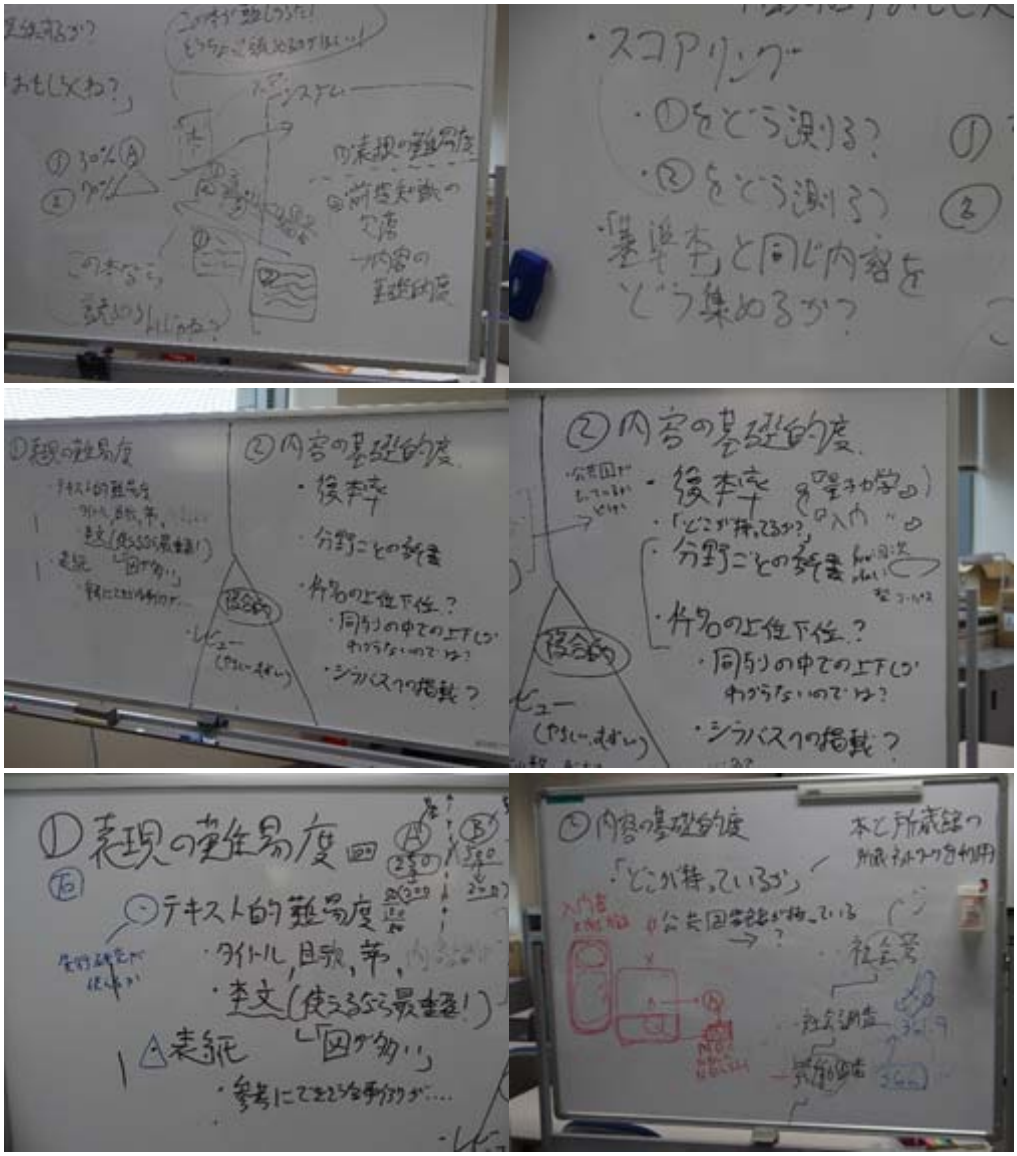


図 2.3-5 システム設計　初期案

（質問・意見）

・推薦というと曖昧な言葉である。システム作成者は、クラスタリング・スコアリング・ランキングという手法を意識しないといけない。

・情報推薦には大きく分けて、アイテムそのものの内容を見るやり方と、アイテム外の関係性を見るやり方がある。奇しくも本システムも 2 軸に分かれていて、しかも「内容の基礎的さ」でのみ所蔵情報という関係性が使われているので、そういった分け方も意識すると良いのではないか。

・内容の基礎的さ、文章の易しさという観点もあるが「近さ」もある。

・図書館の人はメタデータ見ただけで簡単そうか難しそうか、見分けられるのか。

それができるなら、クラウドファンディングでデータを集められる。

・最初の入力の際に、資料のタイトルだけでなく、どれくらい難しいかを 10 段階評価で入力してもらう。

### 2.4. 集合研修後（9/2-11/29）

下記の課題を設定し、まず中間期限日 9/25-9/29 までにそれぞれ作業を行った。

・個々のアイデアの検証

－「同じ内容の本」判定

－「表現の難易度」判定

－「内容の基礎的さ」判定

－スコアリング

－必要なデータ集め

・推薦システムとしてのシナリオの再現

－検証のために「読めなかった」役の書籍を 1 冊選び、データの流れのシナリオを再現してみる。

・実際にシステムを試作し、動作させながら機能を検討する。

### 2.5. 第 3 回集合研修（11/30-12/1）

#### 2.5.1. 1 日目討議～2 日目最終発表

作成したシステムで検証した各結果を共有し、システムによる成果と今後の課題について意見を出し合った。

- 最終発表の資料は「最終報告書」を参照

（質問・意見）

・道のりが具体的に見えたとあったが、今回はどこまで到達しているのか、また最終到達点はどこか。

回答：最終到達点は、ユーザニーズに合ったその人が読みたい書籍を提示することであり、今回のワークショップでは、そのための学習データを作れるようになった。

・利用者が入力できるような機能があるとよい（「いいね」ボタン等）

・フィードバックできるのは大学の教授や修士以上の学生ではないか。

回答：利用者のレベルは限定しない方がよい。初心者が読んでみてフィードバックするとタイムラグが大きいのではないか。開発段階でテストユーザを絞り込み、一般公開後誰でもフィードバックできるとよい。

・基準書より少し簡単、ということが大事で、順位の高低で良い悪い、という議論

にはできない。細かいチューニングより、上位 10 件に入っているか等の評価方法もよいかもしれない。

・基準書を入力すると似たような書籍が推薦される、というサービス（Amazon・Webcat Plus 等）は他にもあるので、このシステムと並べてみせると特徴を説明しやすくなる。

・ 初心者は、どれくらい分からなかったのかは答えづらいが、より平易な表現が欲しいのか、より簡単な内容が欲しいのか、は答えることができる。図書館でも、 利用者が正しい要求を必ずしも言うわけではないが、レファレンスサービスが成立している。専門家にはもっと詳しく聞くことができる。

・このシステムを広く公開・発表できる機会があるとよい。自機関での発表やソースコードの公開等。

# 3. 調査報告

## 3.1. 先行研究

書籍の難易度をテーマにした先行研究を調査した。

### 3.1.1. 「難易度及び類似度を用いたコンピューター関連書籍推薦システムの開発」（舟木・黒田, 2014）[1]

http://ci.nii.ac.jp/naid/110009659647

・ひらがな、カタカナ・漢字・英字、記号、名詞や動詞の多さから難易度を推定

### 3.1.2. 「レビュー情報を用いた学術本の難易度推定」（中山・南保・木村, 2012）[2]

http://ci.nii.ac.jp/naid/130001878759

・専門用語の出現を元に推定

・実験対象における、専門用語辞書を作成し使用

### 3.1.3. その他の先行研究

・「学術書籍の難易度を読者ネットワークから推定する試み」（三好・入野, 2010）[3]

http://ci.nii.ac.jp/naid/110008000785

・「学習コンテンツ推薦を目的とした難易度推定アルゴリズムの評価のための正解データ作成（教育工学)」（三好他, 2014）[4]

http://ci.nii.ac.jp/naid/110009862084

## 3.2. 先行サービス

### 3.2.1. Webcat Plus　連想検索

http://webcatplus.nii.ac.jp/#

Webcat Plus は、国立情報学研究所（NII）が提供する無料の情報サービス。

---

[1] 舟木 類佳,黒田 久泰.難易度及び類似度を用いたコンピューター関連書籍推薦システムの開発.研究報告自然言語処理（NL）.一般社団法人情報処理学会.2014,2014-NL-215(8),1-6,

[2] 中山 祐輝,南保 英孝,木村 春彦.レビュー情報を用いた学術本の難易度推定.人工知能学会論文誌. 2012, 27(3),213-222,

[3] 三好 康夫,入野 美弥.学術書籍の難易度を読者ネットワークから推定する試み."電子情報通信学会技術研究報告. ET, 教育工学".2010,110(67),19-24,

[4] 三好 康夫,濱田 一伸,鈴木 一弘,塩田 研一,岡本 竜.学習コンテンツ推薦を目的とした難易度推定アルゴリズムの評価のための正解データ作成（教育工学).電子情報通信学会技術研究報告 = IEICE technical report： 信学技報. 2014, 113(482),161-164,

連想検索とは、文書と文書の言葉の重なり具合をもとに、ある文書（検索条件）に近い文書（検索結果）を探し出す検索技術である。

#### 3.2.2. Amazon

https://www.amazon.co.jp/

通信販売サイト。レコメンデーション機能がある。

### 3.3. システム構築に役立つツール

#### 3.3.1. API

さまざまなAPIについて調査した。

・書誌データ（タイトル・著者名など）：NDLサーチ[5]、openBD[6]、yahoo[7]、楽天[8]

・目次：openBD

・レビュー：Yahoo!、楽天　※レビュー数が少なく、有効な情報を得るのが難しい

・所蔵：CiNii Books[9]、カーリル[10]

- 詳細は別紙1

#### 3.3.2. テキストの難易度を測定

・帯3 日本語テキストの難易度を測る

http://kotoba.nuee.nagoya-u.ac.jp/sc/obi3/

入力された日本語テキストの難易度を測定するシステム。

・jReadability　日本語文章難易度判別システム

https://jreadability.net/ja/

日本語テキストを入力すると、テキストの概要、品詞構成、語種構成、文字種構成が出力されるシステム。

・画像処理による日本語文章の難易度判定システム

http://blog.hotolab.net/entry/stringdepth

---

[5] (http://iss.ndl.go.jp/information/api/)

[6] (https://openbd.jp/)

[7] (https://developer.yahoo.co.jp/webapi/shopping/)

[8] (https://webservice.rakuten.co.jp/api/booksbooksearch/)

[9] (https://support.nii.ac.jp/ja/cib/api/b_opensearch_lib)

[10] (https://calil.jp/doc/api.html)

文章の濃さに着目した日本語文章の難易度判定システム。入力文字列を画像に変換し、文字ごとの濃度を求め、その平均を文章の濃度とする。

### 3.3.3. 機械学習技術(Word2Vec)

Word2Vec は Google 社の Tomas Mikolov 氏らが 2013 年に発表した手法であり、ニューラルネットワークという機械学習の一手法を用いてテキストデータの単語の並び方を学習することで、各単語を、埋め込み表現と呼ばれるベクトル表現（その単語の特徴を表す数値的な表現)に変換する。

Word2Vec には、文章中で互いに近接して出現しやすい単語同士が似たベクトル表現になるという特徴がある。これを用いることによって、単語同士の関係性の近さを、距離のような定量的な表現で比較することができる。

学習方法の違いとして、前後の単語を使ってその間の単語を推測する Continuous Bag-of-Words(CBOW)と、ある単語を使って、その前後の単語を推測する Skip-gram と呼ばれる 2 つのアルゴリズム がある[11]が、今回は前者を採用したため、CBOW についてのみ説明する。

CBOW はある単語の出現を、その単語の前後の単語から推測できるように学習を行う。

例えば、泉鏡花の高野聖の冒頭文「参謀本部編纂の地図をまた繰開いて見るでもなかろう」から「地図」の出現を、前後 3 単語を使って学習するとする。

前の 3 単語である、「参謀本部」「編纂」「の」

後ろの 3 単語である、「を」「また」「繰」

の計 6 単語を情報として与えたときに、間に入る単語が「地図」と推測できるように学習器を訓練する。学習を進めることで、各単語について、それぞれの特徴を表すベクトル表現が副次的に得られる。

---

[11] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013). (https://arxiv.org/pdf/1301.3781.pdf)

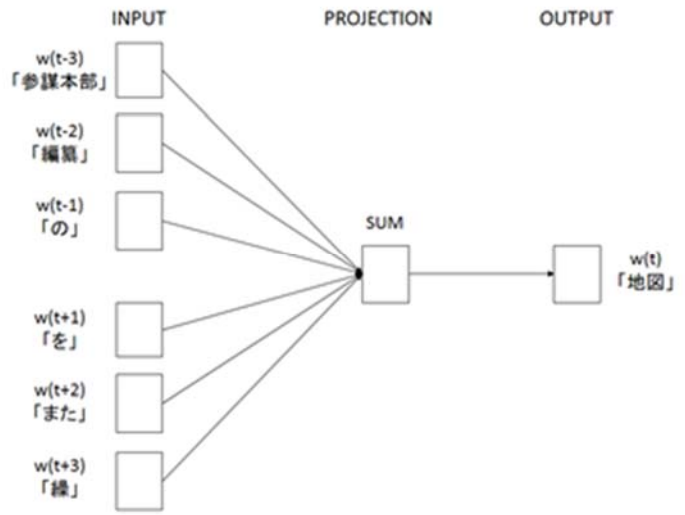



図 3.3 CBOW の概要図(注 11 の原著論文を参考に作成、w は単語の one-hot ベクトル表現)

### 3.4. 入門書とは何か

ある書籍を入門書としたり、2 冊の書籍を比べてどちらが易しいかを推察したりする場合、ヒトは「表現の易しさ」と「内容の基礎的さ」といった観点から総合的に判断している。また、ある分野の学習を始める際、初心者向けの資料をどうやって発見しているだろうか。機械的な処理を行うにあたって、こういったヒトが何気なく行動・判断している方法を明確にする必要があった。

#### 3.4.1. 表現の易しさ

日本語表現としての難易度や表紙のデザイン等を考察する。

・テキスト的な難易度

－「表現の易しさ」にかかわる単語が出現しているかどうか。

- 詳細は別紙 2

－タイトル、目次、帯、内容紹介、（取得可能であれば本文）を対象とする。

・表紙

－対象者層が表紙のデザイン等から推察できるか。

－画像認識（表紙画像 ⇒ 感情判定までなら事例あり）

#### 3.4.2. 内容の基礎的さ

書籍の内容がどれくらい基礎的・入門的であるかを考察する。

・「内容の基礎的さ」にかかわる単語が出現しているかどうか。

- 詳細は別紙 2

・書籍の件名の概念的な上位下位→上位である方が基礎的？
・シラバスに掲載されているかどうか
・複本率の高さ
－ 内容が基礎的な入門書は、より細分化された専門分野を扱う本よりも、
内容が関係する分野が多い
－ 一機関で図書室を複数持つ場合、専門ごとに分かれている
－ したがって、内容が基礎的な入門書は同一機関で複数の図書室に
所蔵されやすく、複本が発生しやすくなると考えられる

### 3.4.3. ヒトはどのようにして入門書を見つけるか？

ヒトが入門書を探す方法として以下が挙げられる。

・Google 検索
－「学習したい分野名」＋「入門」を検索語とする。
・Amazon
－検索して上位に現れるもの
・シラバスでの紹介
－授業シラバスで「教科書」や「参考図書」として示されるもの
・ほかの読者のレビュー
－「入門向け」や「わかりやすい」といった言葉を参考にする。
・本のタイトル
－「入門」等の単語が含まれているもの
・図書館における貸出数・予約数（テスト前など）
・選書カタログにある入門者向け・難易度を示す★を確認する。
・所蔵館（総合図書館 or 専門図書館）の違い
・複本率の高さ
・レファレンス協同データベース
－過去の事例で推薦された書籍を参考にする。

上記を踏まえて、機械的な処理を行う場合の具体的な作業を考えた。

・Google 検索
－「分野名 + 入門」などでヒットするかどうか。
・Amazon 上位
－「分野名＋入門」などでヒットさせた場合の出現順位
・シラバスの紹介
－検索エンジンで取得できる授業シラバスで、示されている本の出現数を取る。
・ほかの読者のレビュー

－「入門」のような単語が文中に出現するか。
－内容がポジティブであるか・評価点が高いか。
・本のタイトル
－「入門」のような単語が出現するか。
・図書館における貸出数・予約数
・所蔵館（総合図書館 or 専門図書館）の違いや複本率の高さを調べる。
・レファレンス協同データベース
－過去に入門向けとして紹介されたことがあるか。

## 4. システムの仕組み

※ユーザが提示する 1 冊の書籍を「基準本」とする。

### 4.1. 対象とするユーザ

システムが対象とするのは、ある本を読みたかったが、内容の専門性の高さ、表現の難しさ等を理由に理解することができなかったユーザとする。

### 4.2. 「相対的入門書」とは

同内容を持つ書籍同士を比べて、より内容が基礎的で、文章が易しく読みやすいものをいう。ユーザそれぞれの状況における相対的な入門書、つまり、読むことで理解できなかった点を補い、当初の目的だった書籍（基準本）の読書へと繋げることが期待されるものである。

本システムでは、基準本と比べると少しだけ入門さが勝るものから順番に出力している。

### 4.3. 「同じ内容」の本を収集・抽出：NDL サーチ API

既存の書籍の集団の中から、ユーザが提示した基準本と同じ内容の書籍のグループを判別し、この中から「入門書らしさ」の高い書籍を抽出することで、基準本と同じ内容でかつ、より入門書らしい書籍を推薦する。

事前に NDL サーチ API から NDL が所蔵する書籍のメタデータを収集し、収集した書籍の ISBN を使用して openBD からより詳細な内容のテキストデータを取得する。

この内容テキストデータを使用して、基準本および NDL から取得した書籍について Word2Vec によるベクトル表現を構築して基準本と各書籍との類似度を測り、類似度の高い本の上位 50~150 件を「同じ内容の本」として採用する。

### 4.4. 「入門書らしさ」のスコアリング

3.4 節での考察から、本システムの「入門書らしさ」は、「表現の難易度」と「内容の基礎的さ」の独立した二軸によって取り扱うことを目指す。

#### 4.4.1. 「表現の難易度」のスコアリング

「表現の難易度」としては、openBD から取得した本のテキストデータについて、以下の要素を扱う。

ただし、ワークショップ第 2 回で指摘された情報推薦における内容ベース・関係ベースの分け方を意識し、出現単語における「表現の易しさに関わる単語」「内容の基礎的さに関わる単語」は、どちらも「表現の難易度」のスコアリングとして扱う。

3.1.1 の先行研究より、

・「ひらがな率」

取得できたテキストデータにおける、ひらがなの割合。これが高いほど「入門書らしい」とする。

・「名詞率」

取得できたテキストデータにおける、名詞の割合。これが低いほど「入門書らしい」とする。

3.3.2 のテキスト難易度判定事例より

・「黒色率」

取得できたテキストデータを画像化した際の黒色ドットの割合。これが低いほど「入門書らしい」とする。

3.4 の考察より

・「表現の易しさに関わる単語出現数」

・「内容の基礎的さに関わる単語出現数」

3.4 の考察において導出した、「入門書らしい単語」の出現数。これが高いほど「入門書らしい」とする。

これら 5 つの要素の値について、それぞれに「類似した本」の中で 0.0-1.0 の範囲に正規化し、それを合計した値を「表現の難易度」とする。

#### 4.4.2. 「内容の基礎的さ」のスコアリング

2.1.2 節の第 1 回ワークショップで上がった「入門書は複本率が高くなりやすい」との意見、および 3.4.2 での考察から、複本率をもって「内容の基礎的さ」を判断する。
ただし実際に CiNii Books API から複本の取得を試したところ、割合よりも単純な複本件数を取得した方が「入門書らしさ」としては有効と判断したため、複本件数を使用する
これを「類似した本」の中で 0.0-1.0 の範囲に正規化した値を、「内容の基礎的さ」とする。

### 4.5. ユーザへの推薦

2 次元グラフに「表現の難易度」と「内容の基礎的さ」によって「基準本」および「同じ内容の本」50 件をプロットして散布図を作り、この散布図上で「基準本」よりも右上に位置する本、すなわち「表現の難易度」と「内容の基礎的さ」の両方が「基準本」よりも勝る本が、推薦対象となる。

この推薦対象グループを一覧としてユーザに提示するにあたっては、基準本より少しだけ入門書らしさが勝る本から順に並べる。

## 5. 開発したシステムの説明

### 5.1. ハーベスト機能について

Python で実装した。コマンドラインで取得したい NDC を指定して実行する。NDL サーチ API の OAI-PMH を利用して、「タイトル」「著者」「ISBN」「価格」「NDC」を指定された NDC について全件取得している。また、NDL サーチで取得した ISBN を利用することで CiNii Books API から「所蔵館数」「各館の所蔵冊数の合計」を、openBD から「目次内容」をそれぞれ取得している。

### 5.2. 機械学習(Word2Vec)による「同じ内容の本」候補の提示について

#### 5.2.1. 学習データ

Wikipedia 日本語版の記事全文(2017 年 8 月断面)について、日本語形態素解析ソフトウェア(MeCab[12][13])と分かち書き辞書(Neologism dictionary[14])を用いて分かち書きを行ったテキストのうち、名詞のみを学習データとした。

#### 5.2.2. 学習方法

Python の自然言語処理ライブラリ gensim [15]を用いた。
単語のベクトル表現の次元数は 100 次元とし、記事中に最低 4 回出てきた名詞を対象とした。学習対象となる単語の前後 5 単語の名詞との関係を学習した。

#### 5.2.3. 類似図書の提示方法

書籍間の類似度の計算には、対象となる書籍のタイトル・目次情報・内容説明に含まれる名詞から、Word2Vec によって計算されたベクトル表現の内積(コサイン類似度)を用いた。

コサイン類似度の説明について、例えば、ミカンが(0, 0.9, 0.4)、ブドウが(1.0, -0.1, 0.1)、グレープフルーツが(0.1, 1.0, 0.6)でそれぞれ表現されるとすると、

ミカンとブドウの類似度は　　0*1.0+0.9*(-0.1)+0.4*0.1=**-0.05**
ミカンとグレープフルーツの類似度は　　0*0.1+0.9*1.0+0.4*0.6=**1.14**

となるので、ミカンに対する類似度は、ブドウよりもグレープフルーツの方が高い、と評

---

[12] (http://taku910.github.io/mecab/)

[13] Taku Kudo, Kaoru Yamamoto, Yuji Matsumoto: Applying Conditional Random Fields to Japanese Morphological Analysis, Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004), pp.230-237 (2004.)

[14] (https://github.com/neologd/mecab-ipadic-neologd)

[15] (https://github.com/RaRe-Technologies/gensim)

価される。

### 5.3. 「表現の易しさ」と「内容の基礎的さ」のスコアリング評価について

【4.4 システムの仕組み】で議論した、「ひらがな率」「名詞率」「黒色率」「簡単そうな単語出現数」「基礎的な単語出現数」をそれぞれ 0~1 の範囲に正規化したものを組み込んだ。以下では、「表現の易しさ」のスコアを easyPoint、「内容の基礎的さ」のスコアを basicPoint とそれぞれ称する。

### 5.4. 実装した Web サービスの技術仕様について

#### 5.4.1. 仮想サーバについて

Amazon Web Service(AWS) t2.medium インスタンス(CPU 2 基(Intel Xeon), ストレージ 30GB, メモリ 4GB)を用いた。T2 インスタンスは、ハイパフォーマンスな演算を行える時間数が限られる(1 時間当たり 24 分。未使用分は 24 時間までストック可能)代わりに低コスト(1 時間当たり 0.046 ドル)で利用可能という特色がある。

Elastic IP を取得し、お名前.com で取得したドメイン(nii-workshop2017.work)と関連づけた。

#### 5.4.2. バックエンドについて

- 第 2 回の報告会において Word2Vec を用いて同じ内容の本の列挙を実験した際に Python でコーディングした。
- 特徴量作成チームに Python の知識があった。

といった観点から、第二回までの Python のコード断片を活用することを目的に、Web フレームワークに Tornado[16]を採用した。Tornado は FriendFeed 社および Facebook 社が開発したのちにオープンソース化されたフレームワークである。

#### 5.4.3. フロントエンドについて

フロントエンド部分の作成に使用したライブラリは以下の通り。

| | |
|---|---|
| **UI** | Bootstrap, jQuery UI |
| **データ可視化** | Google Chart Tools |
| **テーブル表示** | Columns.js |
| **サーバとの通信** | Ajax |

その他 javascript が要求される部分については jQuery で実装した。

---

[16] (http://www.tornadoweb.org/en/stable/)

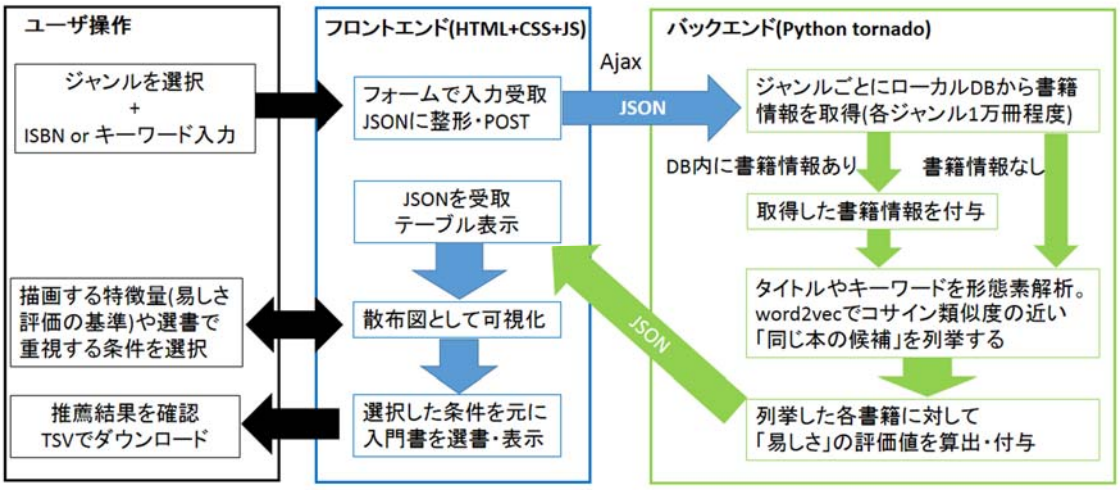



図 5.4.完成したシステムの概要図

## 5.5. Web サービスの利用手順

### 5.5.1. サービスの URL

今回作成したシステムは、2018 年 1 月現在はブラウザから以下の URL にアクセスして使用する。

http://nii-workshop2017.work

### 5.5.2. 初期画面 UI

初期画面では、調べたい分野を指定したうえで、ISBN またはタイトル名等のキーワードで検索を行う。


①日本十進分類による検索対象分野の選択(医学、経済学、社会学の中から選択)[必須]

②一回の検索で提示する「同じ内容の本」の冊数選択(50 冊、100 冊、150 冊の中から選択)[必須]

③ISBN の入力(ISBN-10 または ISBN-13 を半角数字で入力。ハイフンを含んでもよい)[※]

④タイトル名やキーワードの入力(英語または日本語の文字列)[※]
※どちらか必須

### 5.5.3. 「同じ内容の本」の検索結果表示 UI (テーブルによる表示)

検索結果では、まず上段に類似度の高い順に同じ内容の本候補がテーブルとして並ぶ。テーブル内の検索や項目を指定してソートも行える。この画面で書籍の価格や所蔵館数等を得ることができるが、ISBN をクリックすることでカーリルの当該書籍ページにリンクし、より詳細な情報にアクセスすることも可能である。

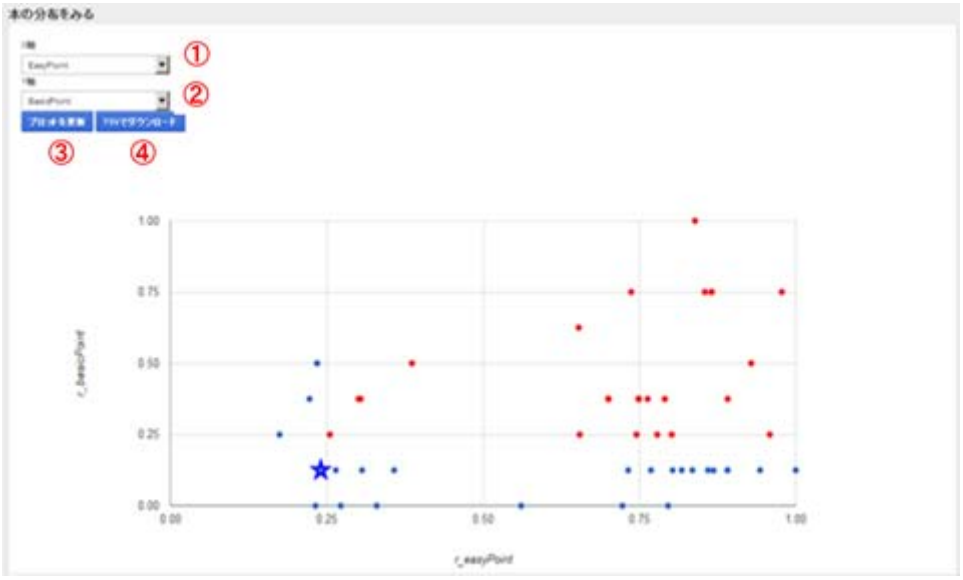


①テーブル内を検索(文字列入力)
②列を指定してソート(列名をクリックする)
③カーリルの該当ページへのリンク(ISBN をクリックする)

### 5.5.4. 「同じ内容の本」の検索結果表示 UI (散布図による表示)

また、検索結果として得られた同じ内容の書籍について、各評価基準によるスコアを散布図で視覚的に把握することができる。散布図中で、検索の元となった書籍は青い★マークで表され、調べたい書籍よりも「入門的である」とされた書籍は赤い点で表される。これらの点はマウスカーソルを重ねることで説明が表示され、クリックするとカーリルの該当

ページにリンクする。

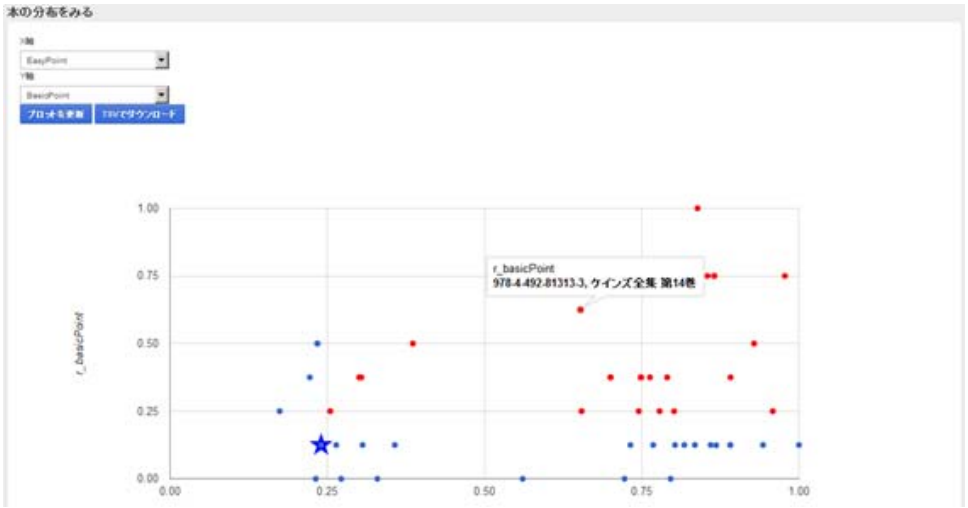



①X 軸とする評価指標の選択

②Y 軸とする評価指標の選択

③選択した指標でプロットを更新するボタン

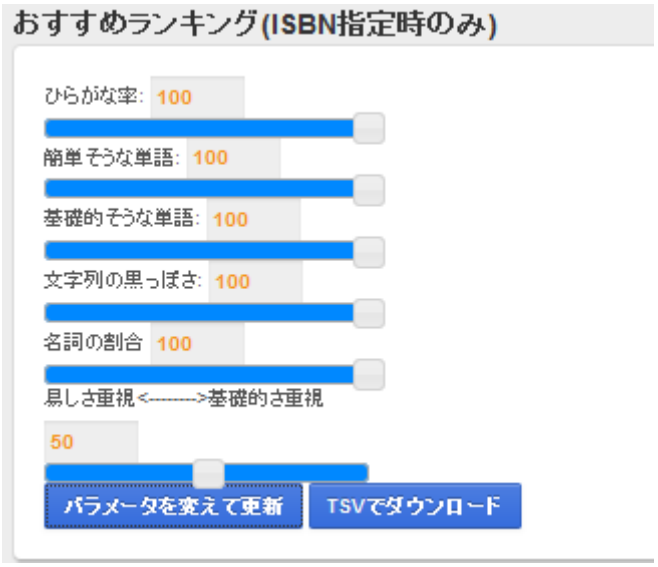
④各種指標を tsv で出力するボタン




散布図内の点にカーソルを重ねたときの表示

### 5.5.5. 「表現の易しさ」と「内容の基礎的さ」のパラメータチューニング UI

散布図の x 軸と y 軸に使われる項目はユーザが変更することができ、散布図を見ながら、パラメータ調整画面でユーザが入門書推薦を利用する際に重視するポイントをチューニングすることができる。




パラメータチューニング後、更新ボタンを押すと、内容に基づいてユーザにあった入門書が、最初に入力した書籍から近しいものから順に、テーブル形式で推薦される。推薦上位 3 件については、Amazon に書影があるものであれば画像が表示される。

# 6. システムの検証

## 6.1. 検証 1　大学の講義ウェブサイトの文献リストによる検証

### 6.1.1. 検証 1-1　「同じ内容の本」の集合がどの程度文献リストと合致するか

大阪大学経済学部のミクロ経済（2015 年 2 学期）が、下記講義ページを設けており、この中の「ミクロ経済学の関連文献」によって、難易度が明記された文献リストが得られる。リストの下に行くほど難易度が高くなる。

https://sites.google.com/site/yosukeyasuda2/home/lecture/micro15

このうち、「ミクロ経済学の代表的・特徴的な教科書（下に行くほど難易度アップ）」として挙げられている 10 件について、それぞれを「基準の本」とした場合に、「同じ内容の本」の集合に他の 9 件がどの程度含まれるかを検証した。「同じ内容の本」は 50 件、100 件、150 件の集合が得られるため、この 3 パターンの設定で検証した。結果は、ウェブサイトの文献リスト順（下にいくほど難易度が高い）で、以下の通りである。

| | タイトル | 著者 | 50 件 | 100 件 | 150 件 |
|---|---|---|---|---|---|
| **書籍 1-1** | ミクロ経済学の第一歩 | 安藤至大 | 1 | 2 | 2 |
| **書籍 1-2** | 今までで一番やさしいミクロ経済学 | 木暮太一 | 0 | 0 | 0 |
| **書籍 1-3** | ミクロ経済学 Expressway | 八田達夫 | 1 | 1 | 2 |
| **書籍 1-4** | ハバード経済学 II 基礎ミクロ編 | グレン・ハバード、アンソニー・ブライエン | 1 | 2 | 2 |
| **書籍 1-5** | **スティグリッツ ミクロ経済学** | **ジョセフ・スティグリッツ** | **3** | **3** | **4** |
| **書籍 1-6** | 経済学 戦略的アプローチ | 梶井厚志、松井彰彦ミクロ | 0 | 1 | 1 |
| **書籍 1-7** | 入門ミクロ経済学 | ハル・ヴァリアン | 2 | 3 | 3 |
| **書籍 1-8** | ミクロ経済学［増補版］ | 武隈慎一 | 0 | 1 | 2 |
| **書籍 1-9** | ミクロ経済学［増補版］ | 林貴志 | 0 | 1 | 1 |
| **書籍 1-10** | ミクロ経済学 | 奥野正寛 | 0 | 0 | 2 |

判別器の「同じ内容の本」の集合が、講義ウェブサイトの文献リストに最も合致したのは、書籍 1-5『スティグリッツ ミクロ経済学』であった。この書籍の結果は 50 件・100 件

の集合で 3/10 件、150 件の集合で 4/10 件であり、150 件の集合で得られた 4 件の内訳は、書籍 1-1、書籍 1-3、書籍 1-4、書籍 1-8 であった。

また、書籍 1-5『スティグリッツ ミクロ経済学』を基準の本とし、150 件の集合で ISBN で検索した結果、計 48 件、ISBN 重複書籍を除いて以下 33 件の書籍が推薦されたが、この 33 件に含まれていたのは、書籍 1-1、および書籍 1-8 であった。書籍 1-1 については判別器の難易度判定と講義ウェブサイトが示す難易度が合致したと言えるが、書籍 1-8 については、逆の結果となった。

#### 6.1.2. 検証 1-2 講義ウェブサイトの文献リストに最適化したパラメータは何か

検証 1-1 では、推薦結果について、講義ウェブサイトと比してあまり高い一致を見ることができなかった。そこで、easyPoint の 5 つのパラメータ（「ひらがな率」「名詞率」「黒色率」「簡単そうな単語出現数」「基礎的な単語出現数」）の重みづけを 25 ポイント刻みで変えながら、検証 1-1 で使用した 10 件の書籍を判別器に入力し、ウェブサイトの文献リストでより簡易であるとされた書籍（より上に記載されている書籍）が、判別器によって推薦されるかどうか、その推薦率を取得した。その結果、講義ウェブサイトの文献リストに最適なパラメータは、以下の通りであった。

ひらがな率　１００
名詞率　７５
黒色率　２５
簡単そうな単語出現数　0
基礎的な単語出現数　0

#### 6.1.3. 検証 1 まとめ

検証 1 では、大学の講義ウェブサイトを基に、パラメータの補正を試みた。これは、ミクロ経済学の特定の講義ウェブサイトの文献リストに対する最適化であり、他の情報があれば、結果が変わることが予想される。いずれにせよ、このような学習データをより多く見つけ、検証することができれば、パラメータ補正の精度向上が期待できる。

### 6.2. 検証 2 実際の研究経験からの検証

#### 6.2.1. 入門書と基準の本

チームメンバーの修士論文執筆経験を基に、システムの検証を行った。執筆した修士論文は、社会学の一分野であるエスノメソドロジーの「ワークの研究」を、図書館のレファレンスサービスで実践したものである。指導教員からは、研究方法について概要を掴むために、何点かの書籍と、類似した研究方法の論文を紹介された。このうち、執筆結果から振り返って、研究方法の概要を理解するための「入門書」として役立った日本語書籍を 2 点挙げる。

【検証用入門書】

書籍２－①　『エスノメソドロジー：人びとの実践から学ぶ』　前田泰樹，水川喜文，岡田光弘 編　新曜社, 2007.8　 ISBN: 9784788510623

書籍２－②　『実践エスノメソドロジー入門』　山崎敬一編　有斐閣，2004.5　ISBN: 978-4641076822

その後、具体的に研究を進める中で参照した書籍のうち、上記２冊よりも難易度の高いものを難易度順に２冊挙げる。書籍２－③は総論的な内容ではあるが、ワークの研究を最初期に進めた研究者の一人による著書で、翻訳書ということもあり、書籍２－①、書籍２－②よりはやや難解である。書籍２－④は、ワークの研究を含むエスノメソドロジーの研究成果をまとめた論文集であり、書籍２－③よりも専門的な書籍であると言える。

【検証用基準の本】

書籍２－③　『エスノメソドロジーと科学実践の社会学』　マイケル・リンチ著　勁草書房　2012.10　ISBN: 9784326602445

書籍２－④　『概念分析の社会学　2　実践の社会的論理』　酒井泰斗ほか編　ナカニシヤ出版 2016.4　ISBN: 9784779510144

上記４冊について、書籍２－③、書籍２－④をそれぞれ基準の本として今回開発したシステムに入力した場合、入門書と見做した書籍２－①、書籍２－②が、出力においてどのように扱われるかを検証した。

### 6.2.2.　検証結果 2-1　書籍２－③を基準の本とした場合

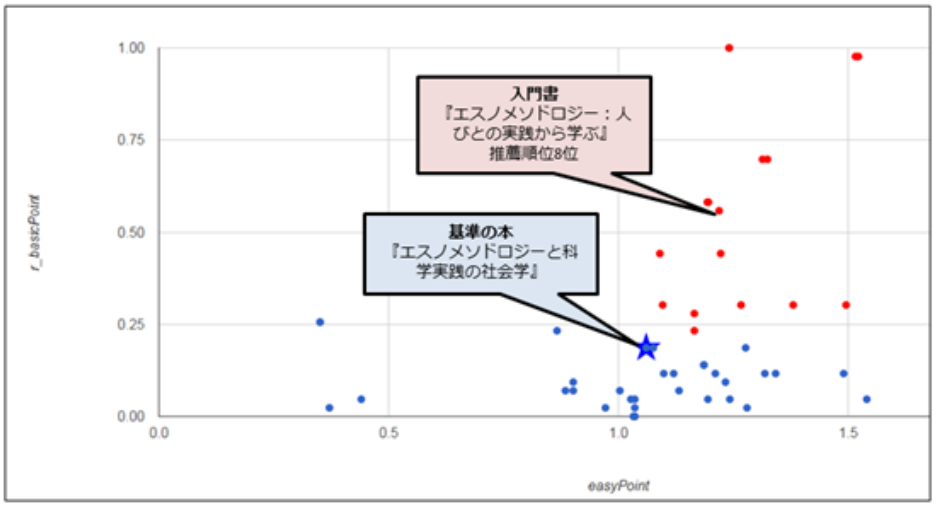
書籍２－③『エスノメソドロジーと科学実践の社会学』を基準の本として ISBN で検索した結果、17 件の書籍が推薦された。実際には同一 ISBN を持つ書籍が重複して出現しており、これを除外すると以下の 13 件となる。

書籍２－③（『エスノメソドロジーと科学実践の社会学』）を基準の本とした際の
推薦結果リスト（同一 ISBN を持つ書籍は重複として除外した）

| 順位 | タイトル | 著者 | rankPoint |
|---|---|---|---|
| 1 | よくわかる社会学 | 宇都宮京子 編 | 0.6988845 |
| 2 | 対人社会心理学重要研究集 7（社会心理学の応用と展開） | 齊藤勇，川名好裕 編 | 0.6993173 |
| 3 | よくわかる社会学 第２版 | 宇都宮京子 編 | 0.7221403 |

| | | | |
|---|---|---|---|
| **4** | コミュニティ心理学ハンドブック | 日本コミュニティ心理学会 編 | 0.7660271 |
| **5** | パラダイムとしての社会情報学 | 伊藤守，西垣通，正村俊之 編 | 0.7845978 |
| **6** | 異文化コミュニケーション・ハンドブック：基礎知識から応用・実践まで | 石井敏 [ほか]編 | 0.8321842 |
| **7** | 臨床社会学のすすめ | 大村英昭，野口裕二 編 | 0.8413152 |
| **8** | 社会構成主義の理論と実践：関係性が現実をつくる | K.J.ガーゲン 著 | 0.8874486 |
| **9** | **エスノメソドロジー：人びとの実践から学ぶ**<br>**【書籍②－１】** | **前田泰樹，水川喜文，岡田光弘 編** | **0.8886368** |
| **10** | 新しい社会学のあゆみ | 新睦人 編 | 0.8988288 |
| **11** | 新版　質的研究入門：〈人間の科学〉のための方法論 | ウヴェ・フリック 著 | 1.0055868 |
| **12** | 構築主義とは何か | 上野千鶴子 編 | 1.1204732 |
| **13** | 質的研究入門：<人間の科学>のための方法論 | ウヴェ・フリック 著 | 1.2463334 |

検索結果のプロットは以下の通りとなる（赤い点が推薦書籍）。




検証の結果は以下の通りである。

1)　「入門書」として想定した書籍のうち、書籍２－①が推薦順位 9 位として推薦された。

2)　「入門書」として想定した書籍のうち、書籍２－②は、推薦されなかった。

3)　推薦順位１位は、社会学の全般的な入門書である『よくわかる社会学（やわらかアカ

デミズム・わかるシリーズ)』（ミネルヴァ書房 ， 2006）であった。この書籍は社会学の入門書であり、「エスノメソドロジー」はもちろん社会学の一概念であるものの、かなり広いテーマである。"「エスノメソドロジー」についてもう少し入門的な書籍が欲しい"という観点からすると、適切な推薦とは言い難い。

4) 「エスノメソドロジー」という用語がタイトルや目次に含まれないにも関わらず、『社会構成主義の理論と実践』（第 8 位）や、『構築主義とは何か』（第 12 位）など、エスノメソドロジーに関連が深い上位概念の概要書が推薦されていた。

### 6.2.3. 検証結果 2-2 書籍２－④を基準の本とした場合

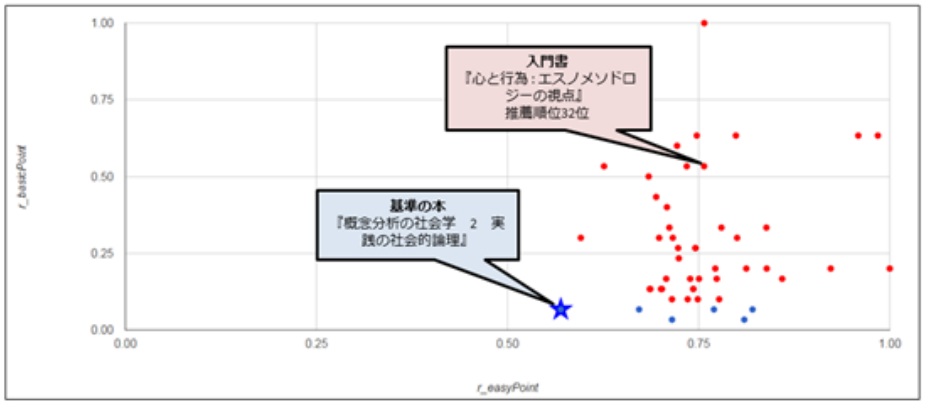
書籍２－④『概念分析の社会学 2 実践の社会的論理』を基準の本として ISBN で検索した結果、計 43 件、ISBN 重複書籍を除いて以下 39 件の書籍が推薦された。

| 順位 | タイトル | 著者 | rankPoint |
|---|---|---|---|
| 1 | 社会心理学 | 白樫三四郎, 外山みどり 編著 | 0.65471253 |
| 2 | パーソナルな関係の社会心理学 | W.イックス, S.ダック 編 | 0.655457006 |
| 3 | 質的研究のデザイン | ウヴェ・フリック 著 | 0.660788436 |
| 4 | トクヴィルとデュルケーム：社会学的人間観と生の意味 | 菊谷和宏 著 | 0.663024556 |
| 5 | 社会学的想像力のために：歴史的特殊性の視点から | 伊奈正人, 中村好孝 著 | 0.667309686 |
| 6 | シリーズ 21 世紀の社会心理学 13 | 高木修 監修 | 0.668473816 |
| 7 | 人間の行動原理に基づいた異文化間コミュニケーション | 西田ひろ子 著 | 0.680643275 |
| 8 | 欲望するシステム | 黒石晋 著 | 0.689870155 |
| 9 | 利他行動を支えるしくみ：「情けは人のためならず」はいかにして成り立つか | 真島理恵 著 | 0.691943764 |
| 10 | 進化するシステム | 中丸麻由子 著 | 0.703611778 |
| 11 | 常識の社会心理：「あたりまえ」は本当にあたりまえか | 卜部敬康, 林理 編著 | 0.71593209 |
| 12 | 考えるヒント：方法としての社会学 | 藤村正之 著 | 0.716794258 |
| 13 | 「日本人らしさ」の発達社会心理学：自己・社会的比較・文化 | 高田利武 著 | 0.726628264 |
| 14 | 階級・ジェンダー・エスニシティ：21 世紀の社会学の視角 | 笹谷春美, 小内透, 吉崎祥司 編著 | 0.737322999 |
| 15 | マックス・ヴェーバーの比較歴史社会学 | スティーヴン・コールバー | 0.746495096 |

| | | | |
|---|---|---|---|
| | | グ 著 | |
| **16** | 社会分析 : 方法と展望 | 金子勇 著 | 0.748616098 |
| **17** | 社会的ジレンマの処方箋 : 都市・交通・環境問題のための心理学 | 藤井聡 著 | 0.753283733 |
| **18** | 現代文化の社会学入門 : テーマと出会う、問いを深める | 小川伸彦, 山泰幸 編著 | 0.761630447 |
| **19** | 社会調査論 | 佐藤健二, 山田一成 編著 | 0.772660801 |
| **20** | 異文化コミュニケーション・入門 | 池田理知子, E.M.クレーマー 著 | 0.776770915 |
| **21** | 人の移動と近代化 : 「日本社会」を読み換える | 中村牧子 著 | 0.796355461 |
| **22** | 岩波講座コミュニケーションの認知科学 = The Cognitive Science of Human Communication 1 | 安西祐一郎, 今井むつみ, 入來篤史, 梅田聡, 片山容一, 亀田達也, 開一夫, 山岸俊男 編集委員 | 0.803478169 |
| **23** | 逸脱とコントロールの社会学 : 社会病理学を超えて | 宝月誠 著 | 0.80738348 |
| **24** | インターネットにおける行動と心理 : バーチャルと現実のはざまで | アダム・N.ジョインソン 著 | 0.811981214 |
| **25** | 人の移動と近代化 | 中村牧子著 | 0.81924277 |
| **26** | 社会学キーコンセプト : 「批判的社会理論」の基礎概念 57 | ニック・クロスリー 著 | 0.819948757 |
| **27** | 地域研究のための GIS | 水島司 | 0.835151478 |
| **28** | 道徳回帰とモダニティ : デュルケームからハバーマス-ルーマンへ | 三上剛史 著 | 0.836063814 |
| **29** | 言説分析の可能性 : 社会学的方法の迷宮から | 佐藤俊樹, 友枝敏雄 編 | 0.836858636 |
| **30** | 社会的認知の心理学 : 社会を描く心のはたらき | 唐沢穣 [ほか]著 | 0.885550272 |
| **31** | ニクラス・ルーマン入門 | クリスティアン・ボルフ著 | 0.891038903 |
| **32** | **心と行為 : エスノメソドロジーの視点 ★エスノメソドロジーに関わる入門書** | **西阪仰 著** | **0.896211961** |
| **33** | 社会調査方法論 | 中道実 著 | 0.91871843 |
| **34** | ニクラス・ルーマン入門 : 社会システム理論とは何か | クリスティアン・ボルフ著 | 0.957253133 |

| | | | |
|---|---|---|---|
| **35** | 偏見の社会心理学 | R.ブラウン 著 | 0.95750331 |
| **36** | 社会 | 市野川容孝 著 | 1.001346287 |
| **37** | 社会学研究法・リアリティの捉え方 | 今田高俊 編 | 1.138627345 |
| **38** | メッセージ分析の技法 | クラウス・クリッペンドルフ著 | 1.149250603 |
| **39** | 社会学研究法 | 今田高俊編 | 1.160602949 |

検索結果のプロットは以下の通りとなる（赤い点が推薦書籍）。




検証の結果は以下の通りである。

1) 書籍２－④は、書籍２－③よりも basicPoint、easyPoint とも低い結果となった。また、抽出された関連書籍の中でも、両ポイントが最も低い結果となった。これは、論文集である書籍２－④は相対的に難しい、という事前想定の通りである。「論文集であるかどうか」を難易度判定の基準にしていないにも関わらず、事前想定と同一結果となった。
2) 「入門書」として想定した書籍２－①、書籍２－②は、いずれも推薦されなかった。
3) 「エスノメソドロジー」を主題とした書籍としては、32 位に『心と行為 ：エスノメソドロジーの視点 』が推薦された。この書籍は概念を解説した概説書であり、著者の経験から判断して、基準の本よりも相対的に難易度が低いと言える。
4) 検証結果２－１同様、基準の本では「エスノメソドロジー」という用語がタイトルや目次に含まれないにも関わらず、32 位でエスノメソドロジーの書籍を推薦している。
5) 推薦順位１位は、社会学の全般的な入門書である『社会心理学』（八千代出版 ，2005）であった。検証結果２－１同様、「エスノメソドロジーを知る」という観点からは、かなり離れたテーマの書籍である。

### 6.2.4. 検証 2 まとめ

検証 2 では、メンバーの実際の研究経験を基に、どのように相対的入門書が推薦される

かを検証した。検証結果 2-1 では、任意の 2 冊のうち 1 冊推薦され、検証結果 2-2 ではメンバーの経験通りの難易度判定がなされた。ここから、相対的難易度の判定は、上手く動いていると言える。

一方で、実際の研究経験から見ると、エスノメソドロジーを研究するという観点から見ると、「同じ内容の本」の集合自体に、関連性が低い書籍が少なからず混じっていた。しかし、これは、エスノメソドロジーという観点が軸になっているためであるとも言え、例えば「社会構成主義」を学ぼうとしている者が、「エスノメソドロジー」ではテーマが具体的過ぎて難しいと考えて本判別器を使用したと想定すれば、より幅広い集合によってこれを補ったとも言える。

この検証結果は、システムの利用者の個別具体的なニーズ、あるいはニーズの背景にある社会的な制約（例えば研究の段階や方向など）によって、こういった判別器の結果も見方が変わってくることを示唆している。

## 6.3. 検証 3　大学図書館の蔵書構成を利用した検証

### 6.3.1. 入門書と基準の本

チームメンバーの 1 人が所属する早稲田大学図書館の蔵書構成を基に、システムの検証を行った。早稲田大学図書館には 21 の図書館・図書室が所在するが、その中でも「中央図書館」と「学生読書室」に着目して検証を行った。

中央図書館の蔵書は「一般図書」と「研究図書」に大きく分けて所蔵されている。このうち一般図書は"主に学部学生を対象とし"たコレクションである[17]。また、早稲田大学は、図書館とは別に、各学部が学部学生用の「学生読書室」を、設置している。これらは学部学生用であるため、各学部の学問分野むけ、かつ比較的学部生向けの難易度の書籍が配架されていると言える。これらの性質を利用し、以下の方法で検証を行う。

(1) 基準の本を選出する。
(2) 基準の本を判別器にかけて、推薦結果の集合を得る。
(3) 各推薦結果の書籍が、①中央図書館で研究図書として何冊所蔵されているか、②中央図書館で一般図書として名冊所蔵されているか、③学生読書室に何冊所蔵されているか、の 3 点を確認する。
(4) 所蔵結果によって、次の 4 タイプに分類する。アルファベットが Z に近づくほど、研究書としての難易度が低いと考えられる。

タイプ A：研究図書のみ
タイプ B：研究図書＋および一般図書または学生読書室
タイプ C：一般図書のみ

---

[17] 早稲田大学図書館ウェブサイト「中央図書館　所蔵資料の配架場所」
(http://www.wul.waseda.ac.jp/CLIB/generalbook.html)

タイプ D：一般図書および学生読書室
タイプ E：学生読書室のみ

本判定器では、推薦順位 1 位の本は、「基準の本よりも少し難しい本」であり、推薦順位が下がるにつれて、基準の本よりも相対的難易度が低くなっていく。そのため、例えばタイプ A の書籍を基準の本とした場合、順位が下がるにつれてタイプ A からタイプ E に近づいて行けば、狙い通りの判定が行われていると言えることになる。

なお、早稲田大学図書館における所蔵は、OPAC（WINE）上の表示を基に次のように判定する。

a. 中央図書館で研究図書として所蔵されている書籍：「中央 B1 研究書庫」または「中央 B2 研究書庫」から始まる請求記号を持つ
b. 中央図書館で一般図書として所蔵されている書籍：「中央 2F 一般図書」「中央 3F 一般図書」から始まる請求記号を持つ
c. 学生読書室に所蔵されている書籍：「政経学読」「法学読」「教育学読」「商学読」「社学学読」「国際教養」「日本語教育」のいずれかから始まる請求記号を持つ

### 6.3.2. 検証結果 3-1　社会学分野のタイプ A を基準の本とした場合

**【検証用基準の本】**

**書籍 3**　『グラウンデッド・セオリー：バーニー・グレーザーの哲学・方法・実践』 V.B. マーティン，A. ユンニルド編　ミネルヴァ書房　2017.2　ISBN: 978-4-623-07372-6 ※中央図書館の研究図書のみで所蔵されており、所蔵タイプは「タイプ A」である

書籍 3（タイプ A）を基準の本として ISBN で検索した結果、計 29 件、ISBN 重複書籍を除いて 24 件の書籍が推薦された。早稲田大学図書館における所蔵状況と、所蔵タイプを示した一覧は、下の表の通りとなる。

| 順位 | タイトル | 著者 | rank Point | 研究図書 | 一般図書 | 学生読書室 | 所蔵タイプ |
|---|---|---|---|---|---|---|---|
| 1 | リーダーシップの意味構成：解釈主義的アプローチによる実践理論の探求 | 片岡登 著 | 0.558513 | 1 | 0 | 0 | A |
| 2 | 理論社会学の可能性：客観 | 富永健一 | 0.60 | 1 | 0 | 1 | B |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | 主義から主観主義まで | 編 | 0697 | | | | |
| 3 | 知識社会学と現代：K.マンハイム研究 | 秋元律郎著 | 0.611735 | 1 | 1 | 0 | B |
| 4 | 社会学の古典理論：数理で蘇る巨匠たち | 三隅一人編著 | 0.628114 | 1 | 0 | 0 | A |
| 5 | グループ・ダイナミックス入門：組織と地域を変える実践学 | 杉万俊夫著 | 0.632283 | 1 | 1 | 0 | B |
| 6 | 臨床社会学ならこう考える：生き延びるための理論と実践 | 樫村愛子著 | 0.638791 | 0 | 1 | 0 | C |
| 7 | ハーバーマス | 小牧治，村上隆夫 共著 | 0.641019 | 0 | 1 | 1 | D |
| 8 | ジンメル社会学を学ぶ人のために | 早川洋行，菅野仁 編 | 0.659777 | 0 | 1 | 2 | D |
| 9 | 組織シンボリズム論：論点と方法 | 坂下昭宣著 | 0.678536 | 1 | 0 | 0 | A |
| 10 | 身体技法と社会学的認識 | 倉島哲 著 | 0.68513 | 1 | 0 | 0 | A |
| 11 | 社会にとって趣味とは何か：文化社会学の方法規準 | 北田暁大，解体研 編著 | 0.699585 | 0 | 1 | 1 | D |
| 12 | 臨床社会心理学の進歩：実りあるインタフェースをめざして | R.M.コワルスキ，M.R.リアリー編著 | 0.703259 | 1 | 0 | 0 | A |
| 13 | 道徳意識の社会心理学 | 片瀬一男，高橋征仁，菅原真枝著 | 0.742642 | 0 | 1 | 0 | C |
| 14 | コミュニティ心理学ハンドブック | 日本コミュニティ心理学会 編 | 0.749885 | 1 | 0 | 0 | A |
| 15 | 社会学をいかに学ぶか | 舩橋晴俊著 | 0.763388 | 1 | 1 | 0 | B |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **16** | 社会学の方法 | 新睦人 著 | 0.768148 | 0 | 1 | 2 | D |
| **17** | はじめて学ぶ社会学 : 思想家たちとの対話 | 土井文博, 萩原修子, 嵯峨一郎編 | 0.798271 | 0 | 1 | 2 | D |
| **18** | 21 世紀の学問論にむけて | 吉田民人 | 0.814445 | 1 | 1 | 1 | B |
| **19** | 社会心理学の新しいかたち | 竹村和久編著 | 0.834937 | 0 | 1 | 0 | C |
| **20** | 数理社会学入門 | 数土直紀, 今田高俊編著 | 0.836812 | 0 | 1 | 2 | D |
| **21** | 社会構成主義の理論と実践 | K・J・ガーゲン著 | 0.870378 | 0 | 1 | 0 | D |
| **22** | 新しい社会学のあゆみ | 新睦人 編 | 0.881258 | 0 | 1 | 2 | D |
| **23** | 社会学の方法 : その歴史と構造 | 佐藤俊樹 著 | 0.884001 | 0 | 1 | 2 | D |
| **24** | 質的研究入門 : 〈人間の科学〉のための方法論 | ウヴェ・フリック 著 | 0.991832 | 0 | 1 | 1 | D |

検証結果は以下の通りである。

1) 上位 5 位は A または B で占められている
2) 15 位より下位には A は出現しない
3) 下位 5 位はいずれも D
4) 7 位から 15 位には、A、B、C、D が順不同で混在している。

1）2）および 3）の結果からは、本判別器が判別した難易度が、蔵書構成の難易度に上手く合致しているということができる。一方、4）の結果は、中間の順位においては本判別器が判別した難易度が、蔵書構成の難易度に必ずしも合致していないことが分かる。この結果から、判別器の難易度判別結果は大学図書館が行う蔵書構成の難易度判別結果におおむね合致するものの、特に中間的な順位においては改善の余地を残しているということができる。

### 6.3.3. 検証結果 3-2　経済学分野のタイプ A を基準の本とした場合

**【検証用基準の本】**

**書籍３**　『世紀転換期の経済と経済思潮 ： 市場と金融』　保坂直達著　晃洋書房, 2001.2　ISBN: 4771012245　※中央図書館の研究図書のみで所蔵されており、所蔵タイプは「タイプ A」である

書籍３（タイプ A）を基準の本として ISBN で検索した結果、計 20 件、ISBN 重複書籍を除いて 16 件の書籍が推薦された。早稲田大学図書館における所蔵状況と、所蔵タイプを示した一覧は、下の表の通りとなる。

| 順位 | タイトル | 著者 | rank Point | 研究図書 | 一般図書 | 学生読書室 | 所蔵タイプ |
|---|---|---|---|---|---|---|---|
| 1 | 貨幣の社会学 ： 経済社会学への招待 | 森元孝 著 | 0.715549 | 1 | 1 | 2 | B |
| 2 | 文化の経済学 ： 日本的システムは悪くない | 荒井一博著 | 0.732493 | 0 | 1 | 2 | D |
| 3 | 経済秩序のストラテジー ： ドイツ経済思想史 1750-1950 | キース・トライブ 著 | 0.737644 | 1 | 1 | 1 | B |
| 4 | イノベーションと内生的経済成長 ： グローバル経済における理論分析 | G.M. グロスマン, E. ヘルプマン [著] | 0.738264 | 1 | 1 | 1 | B |
| 5 | バタフライ・エコノミクス ： 複雑系で読み解く社会と経済の動き | ポール・オームロッド著 | 0.73858 | 0 | 1 | 1 | D |
| 6 | 産業組織論と競争政策 | 小西唯雄編 | 0.764602 | 1 | 1 | 1 | B |
| 7 | 経済思想の源流 | 金子光男編著 | 0.769863 | 0 | 1 | 0 | C |
| 8 | 思想としての経済学 | 竹田茂夫著 | 0.812783 | 0 | 1 | 0 | C |
| 9 | 現代「経済学」批判宣言 ： 制度と歴史の経済学のために | ロベール・ボワイエ [著] | 0.846148 | 1 | 1 | 1 | B |
| 10 | 新講経済原論 | 丸山徹 著 | 0.87468 | 1 | 0 | 3 | B |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **11** | 経済の倫理学 | 山脇直司著 | 0.882873 | 0 | 1 | 0 | C |
| **12** | 経済学の歴史：市場経済を読み解く | 中村達也[ほか]著 | 0.899649 | 0 | 1 | 3 | D |
| **13** | 日本経済がわかる経済学 | 菊本義治, 宮本順介, 本田豊, 間宮賢一, 安田俊一, 伊藤国彦, 阿部太郎 著 | 1.037131 | 0 | 1 | 0 | C |
| **14** | 社会経済学：資本主義を知る | 八木紀一郎著 | 1.121787 | 0 | 1 | 2 | D |
| **15** | 比較政治経済学 | 新川敏光[ほか]著 | 1.132579 | 0 | 1 | 4 | D |
| **16** | 文化経済学入門：創造性の探究から都市再生まで | デイヴィッド・スロスビー 著 | 1.213899 | 0 | 1 | 3 | D |

検証結果は以下の通りである。

1) A は出現しない
2) 11 位より下位には C 以下しか出現しない
3) 下位 3 位はいずれも D
4) 1 位から 13 位には、B、C、D が順不同で混在している。

2）および 3）の結果からは、本判別器が判別した難易度が、蔵書構成の難易度に上手く合致しているということができる。一方、A が出現しないという 1）の結果からは、「少し難しいものを推薦する」という判別結果に疑問が残る。判別結果が正しいとすれば基準の本の難易度がちょうどいき値であったという可能性も考えられるが、任意に選んだ結果であり、その可能性は低いと言える。他の可能性として、基準の本が実際は B 相当である、と考えることもできる。4）の結果は、検証 3-1 で中間的な難易度の判別に難があった現象が、今回は上位でも発生しているということができる。

この結果から、判別器の難易度判別結果は、タイプ D 以上の、特に内容が入門的な書籍については、大学図書館が行う蔵書構成の難易度判別結果におおむね合致するものの、上位から中間的な順位においては改善の余地を残しているということができる。

### 6.3.4. 検証 3 まとめ

検証 3 では、判別器の結果が、大学図書館の蔵書構成における難易度をどの程度再現するかを検証した。

検証結果 3-1 では、上位（難しい書籍）と下位（簡易な書籍）について、判別器の結果と大学図書館の蔵書構成の結果がほぼ一致を見た。検証結果 3-2 では、下位下位（簡易な書籍）について同様の結果を得た。ここから、相対的難易度の判定は、特に入門的な書籍について上手く動いている、ということができる。

一方で、検証 3-1（社会学）と検証 3-2（経済学）では、上位（難しい書籍）の判定に差がでた。この 2 例だけでははっきりと結果として明示できないが、例えば本判別器が考える「難易度が高い本」と、実際に大学図書館で「入門的に扱われる本」は、経済学分野と社会学分野ではやや異なる基準である、といった可能性も考えられる。

## 7. まとめと今後の課題について

今回実装した「相対的入門書判別器」は、利用者が与えた「基準本」の情報から、相対的により入門書らしい書籍を提示するシステムの実証を目的として開発された。

この「相対的入門書判別器」は 5 章に示したようなアーキテクチャを持つシステムとして実装され、また 6 章で示したような出力を実現したが、当初の目的から残された課題、および実際に構築する上で新たに判明した課題が多く存在する。

本章ではこの課題について記述する。

### 7.1. 適用可能分野の拡大

5.1 で示したように、今回実装した判別器においては、事前にある NDC 分類番号を指定して NDL サーチ API から書籍のメタデータを収集し、このメタデータをもとに CiNii Books API や openBD API による情報収集を行っている。この実装から今回構築したシステムの利用段階にあたっても、「基準本」が属する NDC 分類番号を利用者に指定させ、その分類番号で収集した書籍の中から推薦する形となっている。

しかし、ある「基準本」に適した「入門書」が、必ずしも同じ NDC 分類番号に属するとは限らない。NDC 分類の付与は人間の判断によるものであり一定ではなく、特に科学史などのような複合的な内容の書籍については、近しいテーマに関する図書であっても異なる一つの NDC 番号に分類されてしまうことがある。加えて専門書の難易度は扱う分類の専門性にもよるところがあり、利用ケース次第では専門分野としては関連するが異なった専門分野、すなわち異なる NDC 分類の書籍を推薦することが適切である場合も存在しうる。

このことから本システムは、特定の NDC 分類番号に基づいた書籍を母集団とした推薦ではなく、全ての書籍を母集団として利用できることが望ましいと考えられる。

今回、NDC 分類番号を指定した実装となったのは、NDL に収録されている書籍量が非常に膨大であるためにこれを API によって収集するのに必要な時間と、収集後に扱うために必要なリソースが不足していたことが主な原因であり、これを拡大するためには収集・準備期間と計算リソースを十分に用意する必要がある。また、NDL 全件ほどのデータ量を扱った場合のサービスとしてのレスポンス速度についても考慮する必要があり、この対策には、高次元ベクトル表現に対して時間計算量を低く抑えられる近似最近傍探索ライブラリ(Faiss[18]や NGT[19]（Neighborhood Graph and Tree)等)の導入が一例として考えられる。

---

[18] (https://github.com/facebookresearch/faiss)

[19] Kohei Sugawara, Hayato Kobayashi, and Masajiro Iwasaki, “On Approximately Searching for Similar Word Embeddings”, In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics （ACL 2016）, pages 2265-2275. Association for Computational Linguistics, 2016.

### 7.2. openBD への書籍内容情報収集の依存

5.2.2～5.2.5 で Word2Vec による類似図書の判別を、5.3 では「表現の易しさ」のスコアリングについて記述したが、これらの処理は 5.1 で記した openBD から取得した「目次内容」のテキストデータと書籍タイトルを利用して行っている。

しかし NDL から取得した ISBN の全てについて openBD API から「目次内容」を取得できるわけではなく、openBD に未収録の書籍および ISBN も多く存在しており、そのような書籍については目次内容を利用できず、NDL サーチ API から取得できた書籍タイトルのみを利用して類似度の計算を行っている。このような書籍は openBD から取得できた場合の「目次内容」よりも明らかに情報量が劣っているため、そのような書籍は類似図書の判別や「表現の易しさ」のスコアリングにおいて精度が落ちてしまい、意図した通りに利用者に提示される可能性が低くなってしまう。

したがって、そのような openBD での収録のみに依存することなく、書籍の「目次内容」を取得することが課題となっている。

この対策としては、他の API による書籍情報取得と併用するほか、『「BOOK」データベース[20]』のような有償のデータベースを利用することが考えられる。

### 7.3. 「類似した本」についての利用者ニーズ

本システムでは書籍の類似度計算に機械学習(AI 技術)の一手法である Word2Vec を採用したことで、単純な出現単語の一致を越えて内容の類似した図書を判別するサービスが実現された。

しかしながらある書籍の内容全体に対して類似することが、必ずしも利用者が要望する「類似した本」に合致するとは限らないものと思われる。たとえば複合的な内容の図書について、その中のある一要素について利用者が興味を持っていた場合は、その一要素について書かれた書籍を判別することがより望ましいと考えられるが、本システムは複合的な内容の中から一部の要素のみを利用して類似判定することができない。

この課題から、一冊の「興味があるものの読めなかった本」のみでは、利用者のニーズを捉えるための材料が不足しているのではないか、という問題点が浮き彫りになった。この問題の解決にはより多くの情報を利用者から受け取る必要があると考えられる。

ワークショップ中に発案された解決策としては、一冊ではなく二冊以上の「興味があるものの読めなかった本」を利用者に提示させ、それらに共通する要素を抽出して類似判定に用いることができれば、本課題の対策になるのではないか、といった意見が挙げられている。

---

[20] (http://www.nichigai.co.jp/dcs/index3.html)

### 7.4. 「入門書らしさ」の最適化

本システムでは「入門書らしさ」について、「表現の易しさ」と「内容の基礎的さ」という独立的な二つの尺度によるものと捉えている。しかし、この「入門書らしさ」の算出についてはいまだ試験的な実装の段階にあり、十分な検証と最適化ができているとは言い難い。

「内容の基礎的さ」については、基礎的な内容の書籍ほど関連する研究分野が多く一機関内でも複数図書館が導入するといった考えから、「複本件数」を 0-1 の範囲で正規化した値を使用している。これは書籍の内容そのものではなく書籍と導入館との関係から「内容の基礎的さ」を測ろうとする試みであるが、しかし現在は CiNii Books から取得できた所蔵館と書籍の関係を単純な「件数」にして使用しているため、より適切な解析による精度向上の余地が残されている。

また「表現の易しさ」については「ひらがな率」「名詞率」「黒色率」「簡単そうな単語出現数」「基礎的な単語出現数」の 5 つの要素をそれぞれに正規化した値を算出しているが、しかしこれら 5 つの要素は必ずしも等価に扱うことが正しいと限らない。それぞれの要素の重要度に適したパラメータで重み付けを行うことによって、「表現の易しさ」をより適切に数値化できる可能性があると考えられたため、パラメータチューニングについても検証を行った。

これら「入門書らしさ」の最適化については、入門書とされる複数の書籍について「入門書らしさ」を序列づけたリストを用意し、これを正解セットとして「入門書らしさ」の調整とテストを繰り替えし行う必要があるものと考えられる。

しかしこのような正解セットとして使用できそうな「入門書らしさに基づく序列リスト」についてはワークショップ中での調査ではほとんど見つけられず、少数発見できた事例も 6.1.2 のように、ミクロ経済学という特定の専門分野に限ったものであった。

特定の専門分野のデータセットを正解とした最適化が該当の専門分野以外に対しても適当であるとは限らないため、正解データとなる「入門書らしさの序列」を調査・収集する仕組みを併せて設けることが必要と考えられる。

### 7.5. 結果を提示するユーザインタフェース

本システムでは「入門書らしさ」を「表現の易しさ」と「内容の基礎的さ」という独立的な二つの尺度によるものとして捉えたことから、利用者に結果を提示するインタフェースにおいても、2 次元グラフへのプロットによる表現を構築した。

また、その 2 次元グラフによる推薦対象の書籍の集合を基準本と比較して、より「入門書らしさ」の近しい書籍から順に並べ、テーブルとして表示するインタフェースを構築した。

さらには本システムにより計算された書籍の各数値を TSV データとしてダウンロードし、利用者の側での検証や 2 次利用を助けられるインタフェースも用意した。

このように、今回用意された 3 つのインタフェースはそれぞれに異なる目的によって用意されたものであるが、それぞれの狙いが利用者に対し直観的に伝えられているかについては、課題が残されている。

特に、評価指標でソートされたテーブルを表示するインタフェースについては、「当初の目的だった書籍の読書へとつなげる」という目的から「基準本に入門書らしさが近しい」ものからの順序で並べているが、このテーブルと 2 次元グラフ上に点と配色で表現されたプロットとの関係を把握することは、少なくとも初見の利用者には困難と考えられる。

対策案としては、グラフ上のプロットされた推薦書籍について、単なる赤い点ではなくテーブルのソートに用いられた評価指標の値を表示させるなどが考えられるが、あくまでも推薦結果を利用者へ誤解や支障のないよう円滑に伝達することが目的であるため、考察のみならず実際にインタフェースを試作し、利用者からのフィードバックによる改善を繰り返すことが重要となるだろう。

### 7.6. まとめ

「AI 技術の理解とサービス・業務への適用」がテーマとなった本ワークショップは、大学図書館の書籍の中でも特殊な需要の存在する、「入門書」と呼ばれるような書籍を AI 技術によって判別し、図書館利用者や図書館員に推薦することを目的として始められた。このワークショップで実際に調査しシステム設計を考察する中で、本システムで扱う「入門書」とは「初心者」が自身の理解度では内容が理解できないような書籍に直面した際に、「初心者」と「理解できない書籍」との間のギャップを埋めることができるような本であると整理し、この考えを元に「内容が理解できない書籍」の情報から、その書籍よりも相対的により入門書らしい書籍を提示できるような、「相対的入門書判別器」の提案と実践を行うに至った。

この実践のフィードバックとして得られたものが 7.1~7.5 に挙げた課題であり、いずれも容易に解決の見込みのある課題ではない。しかしながらこれらは今回の実践以前の段階ではいずれも把握すらできていなかった課題であり、これを把握できたことは紛れもなく今回のワークショップによる成果であると言える。「相対的入門書判別器」の実現にあたっては、このような実践と検証を継続していくことが求められており、これもまた残された課題と言えるだろう。

### 7.7. AI 技術を図書館業務に活かすとは

一般的に AI 技術を現実の業務に適用し活用できた事例は、そのほとんどが活用先を細かく限定した上で、その中でチューニングと検証を繰り返す地道な努力によって成り立っている。本ワークショップは「入門書を AI で判別する」といった最初のアイデアから、「ユーザの状況に合った難易度の図書を提示する」という形までアイデアを限定し、実際に動くシステムを試作することでチューニングと検証ができる段階にたどり着いた。逆に言えば、試作したシステムを利用したチューニングと検証から先の段階には大きな労力とコストを要するが、その前提として『活用先を細かく限定』し、実現可能性の高いアイデアにまで具体化し構築していく過程が重要となる。

実現可能なアイデアを構築するためには、データが必要である。

AI 技術とは、人間がこれまで行ってきたような知的な判断をコンピュータに行わせる技術であり、その判断のためのデータを用意することが重要である。特に、AI 技術の中でも機械学習の分野で近年著しく発展しているディープラーニングの活用にあたっては、大量の学習データを用意することが求められる。昨今ディープラーニングを活用した AI 技術導入のアイデアは数多く見られるが、実用化については学習データの用意が大きな課題となっている。

学習に必要なデータを大量に用意するということは、基本的にコストの高い作業である。業務に活かすことを目的とする以上はコストの問題を考慮せざるを得ず、コストに見合う効果が見込めなければ取り組むことができない。しかし『活用先を細かく限定』するのがアイデアの条件である以上、いきなり大きな成果を見込めるアイデアを見出すことは容易ではない。AI 活用のコツは活用範囲を小さく限定することであるのだから、そのような小さな活用に数多く着手し積み重ねていくことが、AI 技術を実際の業務に活用するうえでは現実的な道筋と言える。

そういった『小さな活用』に多く着手するためには、データ取得に必要なコストを小さくすることで、着手することの障壁を下げていく必要があるだろう。ここで、AI 技術の導入においてどのようなデータが必要であるかは AI 技術の活用目的に大きく依存するが、そのようなデータを用意することの難易度とコストの大きさは、必ずしも活用目的に依存するわけではない。二次利用目的で Web 上に一般公開されているデータ、および Web 上で効率的にデータを収集できる Web API が存在するためである。

本ワークショップの成果物には Word2Vec と呼ばれる機械学習を活用しているが、学習に利用したデータは Wikipedia の記事全文データであり、このデータは Wikipedia によって CC-BY-SA および GFDL ライセンスの下に公開されている。また、書籍の情報については NDL サーチ API と openBD によって、所蔵館情報については CiNii Books API により取得することができた。これらのデータを容易に取得することができたからこそ、「相対的入門書判別器」は実現可能なアイデアとしてまとまり、またチューニングと検証の道筋を整えることができたと言える。

その逆に、本ワークショップでは「入門書」の判別方法として「定期テストの直前に貸出頻度が上がる図書」などといった図書館業務の観点からのアイデアが出されていたが、これは貸出データ取得の目途が付かなかったことから実装には至らなかった。容易に取得できるデータが AI 活用のアイデアを生み出す一方で、用意するコストが高いデータはアイデアの実現への障壁となると考えられる。

現在、多くの図書館が Web-OPAC サービスを公開しているが、本ワークショップに取り組む中で、傾向として図書館 OPAC は利用者自身のオペレーションによる利用のみが想定されて構築されることが多く、データの二次利用のための機械的な収集を想定した、十分なレスポンス速度を備えた API を有するサービスは、NDL サーチ API のようなごく少数の事例にとどまっていることが実感された。図書館蔵書検索サイト「カーリル」は、日本国内の図書館の蔵書情報と貸出状況について横断的な検索を実現しているサービスであるが、これは個々の図書館の OPAC をスクレイピングするライブラリを開発することによって成り立っており、実現には高い作業コストが費やされている。またカーリルの公開している検索用 API[21]によって各図書館 OPAC からの情報取得も可能だが、基本的には高速な横断検索を提供するサービスであるため、API の利用にあたり 1000 リクエスト/時の制約があるほか、各図書館 OPAC 側のレスポンス速度がボトルネックとなる場合がある[22]など、各図書館のデータについての大規模かつ継続的な収集においては依然として課題が残されている。

したがって図書館で AI 技術を活用するためには、まず各図書館においてデータセットの公開や API の整備を進めることで、二次利用を前提としたデータ取得が手軽に行える状況を作り、実現可能な『小さな活用』の萌芽を増やしていくことが重要と考えられる。

また、今回のワークショップで開発したシステムは、コンピュータが要約した情報(タイトルや目次内容のベクトル表現への変換、書籍の「易しさ」「基礎的さ」の数値化)を使って、ユーザが大量の書籍やレビューに目を通さずとも、自分の興味と理解力に合った本に出会えることを目指した。人間が直接理解できる情報量や費やせる時間は限られているため、AI 技術を図書館のレファレンス業務に活用し、所蔵する膨大な情報から求める情報を探す手助けをするという方向性は今後も検討する価値があるのではないかと思われる。

---

21 (https://calil.jp/doc/api.html)

22 図書館 API 仕様書｜カーリル(https://calil.jp/doc/api_ref.html)

謝辞
今回のワークショップにてご指導いただいた国立情報学研究所大向一輝先生、ご講演いただいた国立情報学研究所相澤彰子先生、電気通信大学上野友稔様、株式会社カーリル吉本龍司様、研修にてご対応をいただきました国立情報学研究所の皆様に感謝いたします。