

# 相対的入門書推薦器

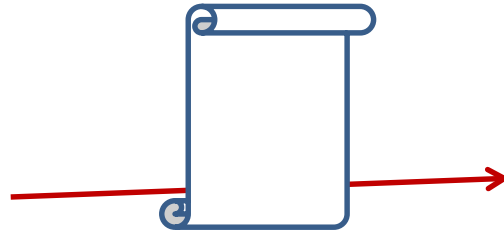
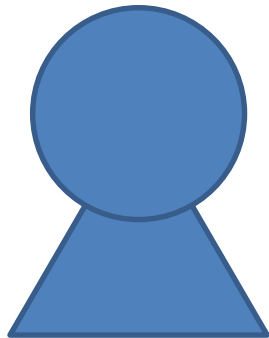
2017年度学術総合ワークショップ

# 相対的入門書推薦器

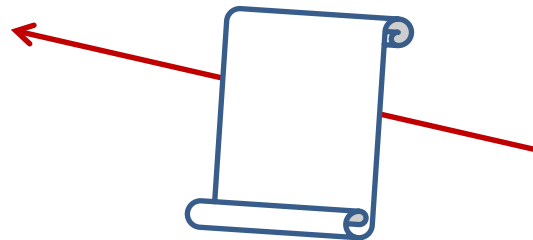
- ユーザが「読んでみたけど読めなかった」本がある
  - 「内容に興味があって手に取ったのに.....」
  - 表現の難易度が高かった
  - 専門的な内容についての基礎知識が足りなかった
- そのユーザが読むべき書籍を提示する
  - 「読めなかった本」と同じ内容で、表現の難易度が低い
  - 「読めなかった本」と同じような主題・内容で、より基礎的な内容について書かれている

# つまりどういうこと？

この本が読みたかったのに  
よくわからなかったんだよ！



相対的入門書推薦器



君は同じテーマについての  
この本を先に読むといいよ！

# 相対的入門書推薦の目的

- ユーザーターゲット
  - ある本を読みたかったが、内容の専門性の高さ、表現のむずかしさから、読むことができなかった人
- 「入門書」とは何か？
  - 上記のユーザでも読める・読むべき本
  - ユーザそれぞれの状況における、相対的な入門書
  - これを読むことで、当初の目的だった読書へと繋げるステップアップが期待できる本

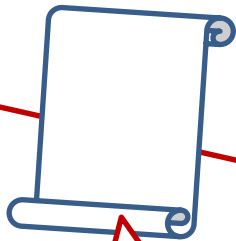
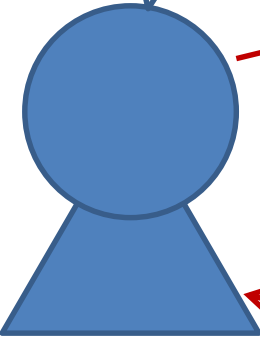
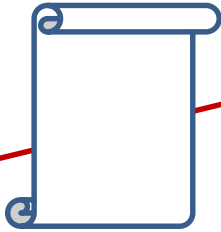
# システムの的に必要なもの

## 相対的入門書推薦器

- ユーザが提示した書籍と、「同じ内容の書籍」の解析
- 書籍ごとの「入門書らしさ」のスコアリング
- 提示書籍よりも「入門書らしさ」が高い書籍の推薦

# つまりどういうこと？

この本が読みたかったのに  
よくわからなかったんだよ！



君は同じテーマについての  
この本を先に読むといいよ！

## 相対的入門書推薦器

「同じ内容」の  
本を収集・抽出

「入門書らしさ」  
の計算

提示された本より  
「入門書らしさ」が高い  
本を判定



# 「入門書らしさ」の計算

- 「ユーザが読めなかった本」と、「同じ内容の本」について、それぞれ「入門書らしさ」を計算する
- 「ユーザが読めなかった本」よりも「入門書らしさ」の高い「同じ内容の本」を推薦する
- 「入門書らしさ」は、どのように計算できるか？

# 「入門書らしさ」

- 表現の難易度
  - 初心者では読めない本が存在するため、まずは読める本から手を出す必要がある
  
- 内容の基礎的さ
  - ある内容について、まずは基礎的な部分を学習する必要がある



# つまりどういうこと？

この本が読みたかったのに  
よくわからなかったんだよ！

## 相対的入門書推薦器

「同じ内容」の  
本を収集・抽出

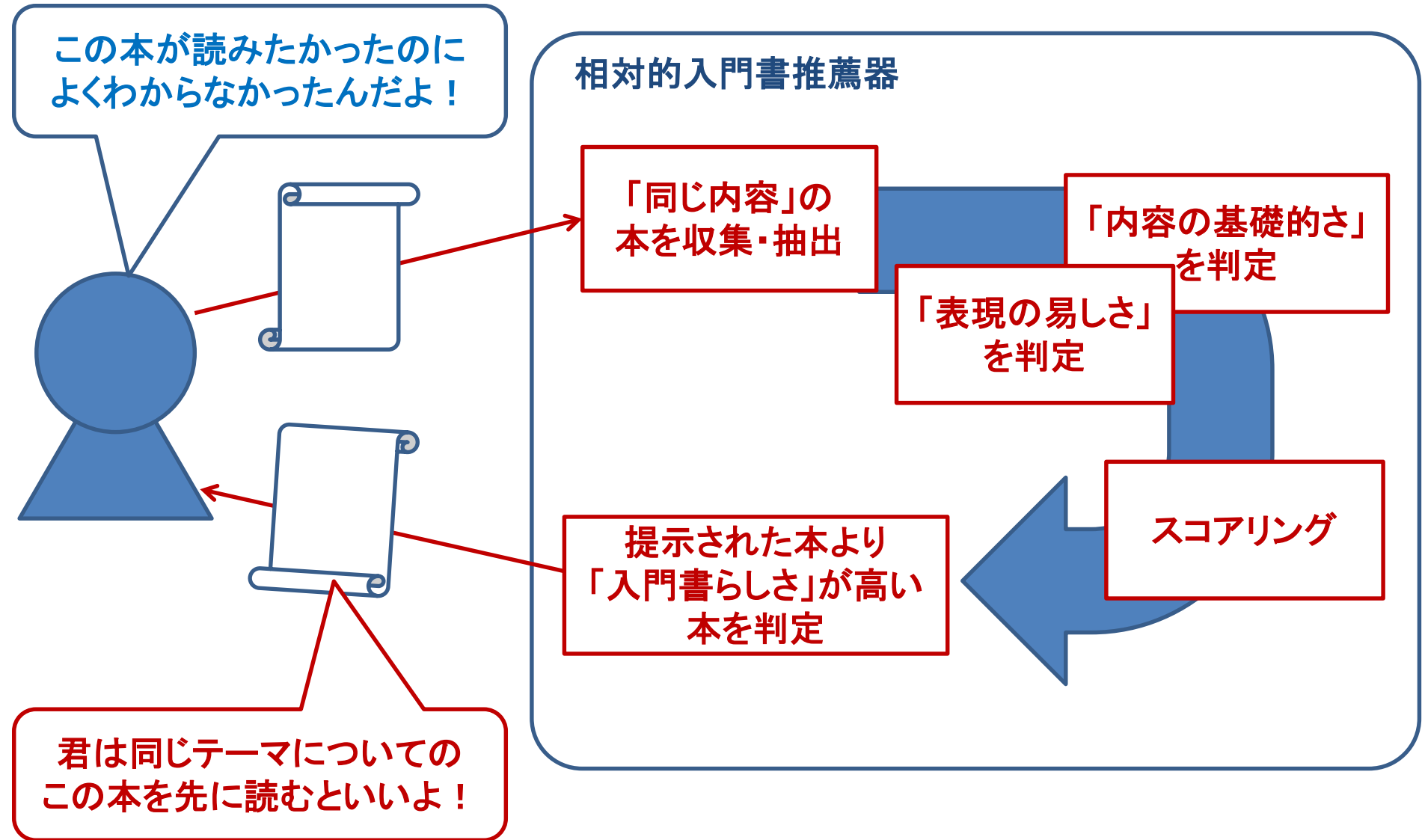
「内容の基礎的さ」  
を判定

「表現の易しさ」  
を判定

スコアリング

提示された本より  
「入門書らしさ」が高い  
本を判定

君は同じテーマについての  
この本を先に読むといいよ！



# 「同じ内容」の本を収集・抽出

- 分類、件名
  - NDCとNDLSHを使って件名から集める
- タイトル・目次で類似度(Word2Vec、Doc2Vec)
  - 機械学習を使って、件名に表れない類似図書を集める

# 入門書らしさ：表現の難易度

- テキスト的な難易度
  - － 「表現の易しさ」に係る単語の出現など
  - － タイトル、目次、帯、内容紹介
  - － 本文（取れるのであれば）
  - － 先行研究あり
- 表紙
  - － 画像認識？
  - － （参考にできそうな事例があまり見つからない  
表紙画像 ⇒ 感情判定までなら事例もあるが.....）

# 入門書らしさ:内容の基礎的さ

- 所蔵館の情報
- 複本率、複本件数
  - 本の内容ではなく、導入実績ベース
  - 多くの機関で必要とされる
    - ⇒ 多くの機関の専門にまたがる基礎的な内容
  - 1機関の中でも複数の館で必要とされる
    - ⇒ 多くの図書室の専門にまたがる基礎的な内容

# 入門書らしさ:内容の基礎的さ

- 「内容の基礎的さ」に係る単語の出現
- 書籍の件名の概念的な上位下位
- シラバス掲載

# 入門書らしさ:複合的

- ユーザレビュー
  - 「やさしい」「むずかしい」などの単語の出現
  - Yahoo, 楽天
  - SNS投稿の解析
- 貸出数、販売・版数

# 「入門書らしさ」のスコアリング

- 「表現の難易度」と「内容の基礎的さ」は独立的であり、同列には比較できない
- どちらの尺度が重要かはユーザに依る部分が多い
- ユーザに提示する形について、2通り検討する
  - 配分の大きさを、ユーザに決めさせる？
  - 2つのスコアによる結果を併記する？

# 現段階の懸念

- 「複本件数」などの各因子は、  
本ごと・分野ごとに、値の多寡の傾向に  
偏りがあるのでは？
  - その場合、スコアリングとして使用する上で、  
こちらが本来想定したバランスが偏りによって  
崩れるのでは？



# 今後の方針

- 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
- 実際に作って動かしてみる