

ニュース記事に関連した書籍推薦システム

森田 真行 浦谷 則好

東京工芸大学工学部コンピュータ応用学科

1. 目的

ニュース記事を読んだときに記事に関連したより詳細な情報を知りたいと思うことがある。コンピュータの普及により、インターネット上に膨大な情報があふれ、検索等で関連する情報を得ることは可能であるが、深く調べるには書籍がまだ一般的なメディアであると思われる。

新聞検索や新聞・雑誌記事横断検索はすでに web 上で提供されている。^{[1][2]} また書籍検索も web 上で提供されているが^{[3][4]}、新聞記事から書籍を検索するものは我々の知るかぎり存在しない。

そこでニュース記事と書籍の双方からキーワードを抽出し、関連付けを行うことで、記事に対して理解や興味などを促してくれる書籍を推薦するシステムを試作した。学習意欲の向上、関連資料の検索時間短縮が補助等を期待できる。

2. システム構築

2. 1 データベース

まずニュース側と書籍側のデータベースを用意する。ニュース側は「CD 毎日新聞データ集 2006」^[5] (以下 N-DB とする)を記事を選択するために利用する。書籍側は「BOOK」^[6]データベース 2006 年度版(以下 B-DB とする)を使用する。

2. 1. 1 N-DB

N-DB は毎日新聞の東京・大阪本社の朝夕刊最終版を対象とした 2006 年の全文記事データ集(タグ付テキストデータ)で、2006 年度版のデータ件数は約 9 万 5000 件である。

2. 1. 2 B-DB

B-DB とは日外アソシエーツ株式会社、他3社が共同構築したもので、図書館やインターネット書店などでも多数導入されている。図書内容を紹介するデータベースであり 1986 年 1 月から収録を開始しており、今も継続中で現在までの蓄積累計総数は約 100 万件である。

データ情報として書誌タイトルのほか概要(帯情報)、目次等のほか、B-DB 独自の書店の棚をイメージした書店分類コード(表 1)や検索用キーワードが各書籍に付与されており内容を掴むことができる。データソースは取次各社の新刊書籍仕入窓口を持ち込まれる新刊委託見本や官庁出版物等が対象であり、収録対象は国内で刊行された書籍や(コミック、雑誌、一部の学習参考書、受験参考書の類、楽譜を除く)資格試験関係の学習参考書である。今回使用したのは 2006 年度

版で、データ件数は約 6 万 3000 件である。

準備として書籍側の B-DB から関連付けに必要と思われる項目と結果表示に必要と思われる項目を抽出し、データベース化しておく。書籍側のキーワードとしては目次、要旨等から抽出されたキーワードの 4 種類で、インデックスを作成しておく。

2. 2 手法

実験の流れを図 1 に示す。まず記事からキーワード抽出を行う。ニュース側は記事見出し文、記事本文を形態素解析システム ChaSen にかけて名詞等をキーワードとして抽出する。この時、掲載面種別コード(表 2)も抽出しておく。掲載面種別コードとは記事の「経済面」、「生活面」など記事が何面に掲載されたかを数字で示したものであり、書店分類コードとの関連付けに使用する。

上記で集めたニュース側キーワードと書籍側のキーワードを用いて対応付けを行う。

書店分類コードと掲載面種別コードとの対応表(表 3)は手動で作成した。

対応付けの流れとして、ニュース側のキーワードとインデックスファイルとの一致の有無を調べて書籍番号を求める。

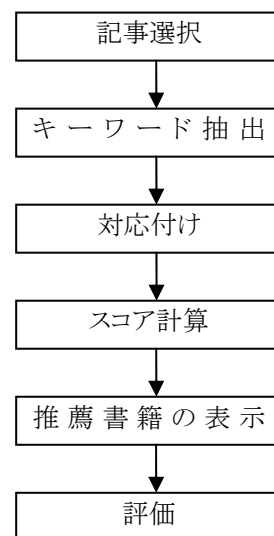


図1 推薦の流れ

表 1 書店分類コード

生活	健康	旅行	趣味	絵本	社会
政治	法律	経済	産業	情報	
テクノロジー	サイエンス	メディカル			
哲学	歴史	語学	教育	芸術	文学
小説(日本、外国、ポルノ)					

表 2 掲載面種別コード

1 面	2 面	3 面	解説	社説	国際	経済
特集	総合	家庭	文化	読書	科学	
芸能	スポーツ	社会				

表3 対応表(一部)

		書店分類コード				
		健康	趣味	情報	...	歴史
掲載面種別コード	1面	○	○	○		○
	経済			○		
	家庭	○	○	○		
	文化			○		○
	科学	○				
	⋮					
	スポーツ	○	○			

表4 書籍インデックス例

キーワード	スコア	整理番号		
あいさつ	3	001	002	005
カレー	2	001	002	
⋮				
学校	1	001		

キーワードのスコアは以下の式で定義した。^[7]

$$\text{idf} = \log_2 (N / \text{df}) \quad (1)$$

N: B-DB に収録された書籍総数

df: キーワードをもつ書籍数

最終的に書籍がもつスコアは、キーワードのスコアと記事中にそのキーワードが出現した回数の積の和で決定される。

$$\text{Score} = \sum \text{tf} * \text{idf} \quad (2)$$

tf: 記事中に出現した各キーワードの回数

式(2)で求めたスコアが各書籍番号に与えられ、高いスコアを持った書籍情報が B-DB を参照して提示される。

3. 実験

まず記事選択として N-DB から付与されている記事 ID をプログラムによってランダムに選択する。その記事の掲載面種別コード、記事見出し文、記事本文を抽出し、提案した手法で推薦書籍を表示する。実験例を下記に示す。

抽出された記事

麻生太郎外相は9日、自民党総裁選へ....正式な出馬表明を行う方向で検討している。

表5 キーワード抽出例(一部)

キーワード	df	idf	tf*idf	個数	番号	
総裁	26	11.26	22.51	2	587	452
麻生太郎	1	15.96	15.96	1	250	
谷垣禎一	1	15.96	15.96	1	250	
安倍晋三	20	11.64	11.64	1	250	452
外相	18	11.79	11.79	1	587	166
財務	238	8.06	8.06	1	166	

表6 スコアリング結果(一部)

書籍番号	スコア	書名
250	43.56	誰が日本を救うのかーポスト小泉・有力政治家に見る人間力
587	34.30	決定版 日本のいちばん長い日
452	34.15	安倍晋三の経済政策を読む

評価は上記の記事の見出し文、記事本文と推薦された書籍を比較し、5段階で評価する。書籍側の情報に書名、要旨、目次から判断してもらう。現在、評価実験を実施中であり、結果は別途報告する。

4. まとめ

書籍推薦システムとしては妥当なものが構築できたがまだ表示のさせ方、推薦順位の付け方に問題が残っている。

今後は被験者数を増やし、より多くの評価実験を行いたい。またスコア付けやキーワード抽出の精度をあげ、さらに対応付けの手法を改善することによってシステムの改良を図りたい。

参考文献

- [1] 新聞記事横断検索 yahoo
<http://gsearch.news.yahoo.co.jp/gs?ty=g>
- [2] 新聞・雑誌記事横断検索 @nifty
<http://www.nifty.com/RXCN/>
- [3] 書籍検索サイト google
<http://books.google.co.jp/>
- [4] 書籍検索サイト 楽天
<http://books.rakuten.co.jp/RBOOKS/>
- [5] CD-毎日新聞データ集
<http://www.nichigai.co.jp/sales/mainichi/mainichi-data.html>
- [6] 「BOOK」データベース
http://www.nichigai.co.jp/newhp/dcs/book_f.html
- [7] 北 研二ほか:情報検索アルゴリズム, 共立出版(2002)