発表タイトル	著者プロファイルを利用した日本全国書誌の自動分類
発表者所属名	日本文学研究専攻
発表者氏名	野本忠司

発表内容

目的

本研究では、書誌レコードの自動分類手法について検討する。書誌レコードの分類は現在すべて人手で行われており、その手間は膨大である。例えば、国立国会図書館の１８年度新規受け入れ図書数は238,107冊でそのうち実際にデータ化された図書数は194,419冊と報告されている （データ化率 81. 7%) (平成１８年度国立国会図書館年報)

手法

本研究では、書誌レコードとして以下のようなものを想定する。

ID	書誌レコード
1	「あいまい」の知 / 河合隼雄,中沢新一編 -- 岩波書店, 2003.3
2	いよよ華やぐ. 上巻 /瀬戸内寂聴著 -- 新潮社, 2001.10
3	国語教師のパソコン/伊井春樹編 -- エデュカ,1989.2
4	イカの哲学 / 中沢新一,波多野一郎著 -- 集英社, 2008.2
5	哲学教室. 第1部 / ヴァージリアス・ファーム編, 植田清次訳編 -- 理想社, 1955.

本研究の関心はこの貧弱な情報を使っていかに正確に書籍の内容に関する分類を行うかという点にある。本件では、この問題に対応するため著者プロファイルという概念を導入する。

本件では、最終的に機械学習法(SVM)とプロファイルの混合モデルを用いて、未分類レコードを自動的に日本十進（最上位）に分類することを試みる。（本手法ではレコード内に出現する最大20文字までの文字列をすべて比較のため用いる。）以下はモデルの詳細。

$$H(\mathbf{x}) = \underset{y \in \mathcal{Y}}{\operatorname{argmax}} \{\alpha f(y, \mathcal{A}^-(\mathbf{x}), h) + (1-\alpha) g(y, \mathcal{A}^+(\mathbf{x}))\}$$

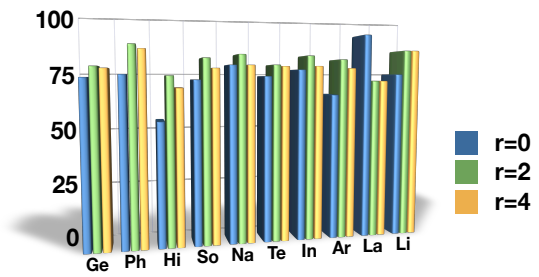
$$f(y, x, h) = P(y = 1 \mid h(x)) = \frac{1}{1 + e^{Ah(x)+B}}$$

$$k(x, x') = \sum_{s \in A^\star} num_s(x) num_s(x') \lambda_s$$

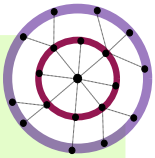
$$g(y, \mathtt{C}) = \text{freq\_ratio}(y) \text{ in } N^r(\mathtt{C})$$

結果

実験の結果、プロファイル情報が書誌分類上、極めて有効であることが確認された。右表参照。その一方で半径をあまり広げると著者の特徴がぼやけ、精度が低下することも確認された。



著者プロファイルとは

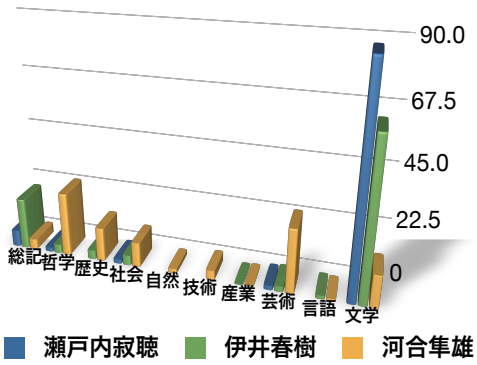


日本十進分類による著者の数値的表現

半径１：著者とその共著者の著作の十進分類の分布
半径２：著者とその共著者、その共著者の共著者の著作の十進分類の分布
半径３：著者とその共著者、その共著者の共著者、さらにその共著者の共著者の著作の十進分類の分布

本研究では、国会図書館OPACからプロファイルを自動構成した。

著者プロファイル



実験

データ: 日本全国書誌 (2006年度)
学習: 1,600 レコード
テスト:1,000 レコード
分類法：日本十進分類（最上位階層）
評価尺度: 幾何平均

以下、GEN （総記）PHIL (哲学）HIST （歴史）SOC (社会）NAT（自然）TECH（技術）IND（産業）ART（芸術）LANG（言語）LIT（文学）

	$r=0$	$r=2$	$r=4$	*freq.*
GEN.	73.8	**78.4**	77.6	2.7
PHIL.	75.0	**87.6**	85.7	4.5
HIST.	55.0	**74.5**	69.2	4.9
SOC.	72.7	**82.1**	77.7	16.4
NAT.	79.1	**83.7**	79.1	6.9
TECH.	74.2	**79.2**	78.6	7.3
IND.	77.1	**83.3**	78.7	4.5
ART	65.8	**81.7**	77.8	18.0
LANG.	**92.9**	72.0	72.0	1.3
LIT.	75.0	**87.8**	85.9	33.5