6M-2

NDCを用いた人物ディレクトリの開発

浦　芳伸†　　村上　晴美†

大阪市立大学大学院創造都市研究科†

1. はじめに

Web 上で情報発信する人々の増加に伴い，Web 上の人物検索は一般的になっており，人物の識別や理解の支援は重要な研究課題となってきている．

Web 検索はキーワード検索とカテゴリ検索に大別できる．Web 上の人物検索ではキーワード検索が主流であるが，キーワード検索だけでは目的の人物にたどりつけないことや，検索された人物の識別や理解がしにくいことがあり，人物を特徴付けるラベル付けが重要であると考える．一方，カテゴリ検索が行える人物検索サービスは存在するがタグ付けによる分類が多く，分類体系を使っている物は見つからなかった．これらの背景から，汎用的な分類体系の一つとして図書館で利用されている NDC9 の分類を人物に付与できれば，人物の識別と理解に有用であると同時に NDC9 の体系でカテゴリ検索ができる人物ディレクトリを構築できると考えた．

本研究では，NDC9 の分類記号を人物に付与する手法を提案し，同手法を用いた人物ディレクトリを開発することを目的とする．

本稿の構成は以下のとおりである．2 節で NDC について説明する．3 節で提案手法，4 節で提案手法の実験について述べる．5 節で試作した人物ディレクトリを示す．6 節では関連研究と比較する．

2. NDC

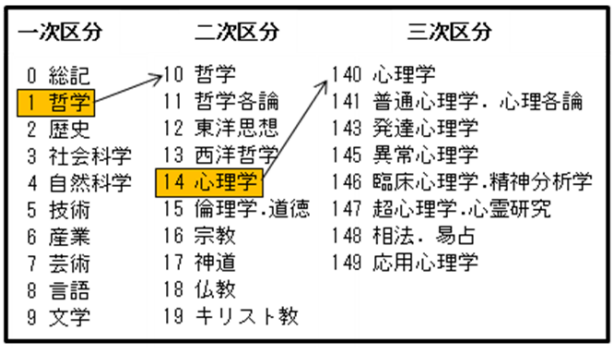

図 1:NDC 三次区分までの例

NDC(Nippon Decimal Classification，日本十進分類法)は，十進方式を導入した，日本で最も一般的な図書の分類法である．1929年に考案され，以後改訂を繰り返している．NDC9は1995年に刊行された新訂9版を指す．最大10分野に繰り返し区切る階層構造を持っている．図1に三次区分までの例を示す．

3. 提案手法

提案手法の主要なアイデアは，NDC9 の相関索引を利用することにある．相関索引は索引語とそれに対応する分類記号の配列である．索引語には分類の細目表に示される名辞をはじめとして必要と判断された用語が含まれている．

NDC9 の MRDF 版の相関索引には，分類記号 8,551 件に対して，索引語は 29,514 件存在する．提案手法は，(1) 相関索引語の抽出，(2) NDC の付与，の 2 段階で構成される．

3.1 相関索引語の抽出

(1) 抽出

HTML のタグの除去後，相関索引語を抽出する．文字列から複数の索引語が抽出できる場合，最も文字数の多い相関索引語を抽出する．

(2) 不要語の除去

一字の語と，Web文書で出現頻度が高いもの100語程度を不要語として除去する．

3.2 NDC の付与

(1) 相関索引語を分類記号に変換

相関索引語を分類記号に変換する．

(2) 関連度計算

以下のとおり人物毎に分類記号 *ndc* のスコアを相対頻度として算出する．

$$score(ndc_i) = \frac{freq(ndc_i)}{\sum_{k=1}^{n} freq(ndc_k)}$$

ただし，n は人物毎の分類記号の異なり数とする．

(3) 解の付与

一次，二次，三次区分の場合は，それぞれ 100 の位，100 と 10 の位，100 と 10 と 1 の位が一致する分類記号のスコアの和の最大値をとるものを解とする．

全区分の場合は，四次区分以下が一致する分類記号のスコアの和の最大のもの，なければ同

Developing NDC-based People-search Directory
† Yoshinobu Ura, Harumi Murakami, Graduate School for Creative Cities, Osaka City University

様にして三次区分以下が一致する分類記号のスコアの和の最大のもの，と繰り返して解を求める．

3.3 例

江川卓氏に関する文書として「江川卓は元プロ野球選手で野球解説者．ワイン好き．」があるとする．「元」「プロ野球」「野球」「ワイン」が抽出された後「元」が除去される．「プロ野球」「野球」「ワイン」はそれぞれ 783.7, 783.7, 588.58 に割り当てられる．783.7 のスコアは 0.666 (2/3)，588.58 のスコアは 0.333(1/3)となり，最大値の 783.7（野球）が江川卓氏の分類記号（全区分）となる．

4. 実験

Yahoo!Web 検索 API より，20 の氏名[1]をクエリとして上位 100 件の HTML ファイルを取得し，人手で同姓同名人物毎に分類されたデータセットを使用する．人物 137 人に対して，一次，二次，三次，全区分の計 4 種類の正解を人手で付与した．

文書は，データから抽出した各文書(タイトル，スニペット，HTML 全文，KWIC 文書)を対象とする．KWIC 文書は人名の前後の文字を抽出した物である．文字の範囲は 50,100,200 文字の 3 種類を使う．

比較手法として，分類項目名（小項目名含む）と文書に形態素解析を行い，tf-idf と余弦を用いた手法を用意した．

上記の各文書，手法，区分で正解率を算出して評価を行った（表 1 参照）．

$$\text{正解率} = \frac{\text{付与された分類記号が正しい人物数}}{\text{人物数}}$$

ただし，正しい分類記号がない人物に対しては「なし」を正解とする．

表 1：実験結果

手法	文書	一次	二次	三次	全
tf･idf	最高値	0.438	0.336	0.226	0.146
cos	最高値	0.431	0.343	0.248	0.168
提案手法	タイトル	0.511	0.453	0.358	0.248
	スニペット	0.445	0.358	0.270	0.204
	全文	0.518	0.431	0.285	0.117
	KWIC50	0.518	0.416	0.292	0.234
	KWIC100	0.511	0.372	0.277	0.197
	KWIC200	0.474	0.365	0.263	0.161

全体的に提案手法の評価が比較手法より高いことがわかる．

一次区分では全文または KWIC50 を用いた提案手法，二次，三次，全区分ではタイトルを用いた提案手法が最も良かった．

ページ数が多い人物では，KWIC 文書（特に KWIC50）を用いた提案手法が良かった．

5. 人物ディレクトリ

提案手法とデータセットを用いて人物ディレクトリを試作した．全文を用いた提案手法により，上位5件の分類記号を人物に付与してから，二次区分以降のカテゴリに割り当てている．

図2に二次区分「78（スポーツ．体育）」の画面例を示す．三次区分の一覧と，二次区分に含まれる人物の一覧が表示されている．

図2:人物ディレクトリ

6. 関連研究

同姓同名人物を識別するために，人物毎に分類されたクラスタに適切なラベルを付与する研究が行われている．Wanら[2]はクラスタ内の文書から抽出した肩書を付与している．上田ら[3]は職業関連情報のラベルを付与している．本研究ではクラスタに関連するNDCを付与している．

清田[4]はキーワードに関連するNDCを提示するが相関索引は利用していない．

7. おわりに

人物を特徴づける NDC を付与するために相関索引に着目した手法を提案した．Web 人名検索結果に適用した実験の結果，本手法の有効性を確認した．提案手法を用いた人物ディレクトリを試作した．

参考文献

[1] 佐藤進也，風間一洋，福田健介，村上健一郎：実世界指向Webマイニングによる同姓同名人物の分離，情報処理学会論文誌：データベース，Vol.46，No.8 (TOD26)，pp.26-36 (2005)．

[2] Wan, X., Gao, J., Li, M. and Ding, B.: Person Resolution in Person Search Results: WebHawk, *Proceedings of CIKM2005*, pp.163- 170 (2005)．

[3] 上田洋，村上晴美，辰巳昭治：Web上の同姓同名人物識別のための職業関連情報の抽出，システム制御情報学会論文誌，Vol.22，No.6，pp.229-240 (2009)．

[4] 清田陽司，リサーチ・ナビ検索システムの技術，参考書誌研究，Vol.71，pp.33-53 (2009)．