The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

Web 上の人物への BSH の付与
Assigning BSH Headings to People on the Web

下倉 雅行 Masayuki Shimokura
村上 晴美 Harumi Murakami

大阪市立大学大学院創造都市研究科
Graduate School for Creative Cities, Osaka City University

We investigate a method that assigns Basic Subject Headings (BSH) 4th edition to the results of web people searches to help users select and understand people on the web. By assigning BSH headings to people, well-formed keywords can be assigned. In this paper, we examine the following combination of factors: (a) web-page rank, (b) position inside HTML, (c) synonyms, and (d) document frequency. We report our results of experiment using an 80-person dataset.

1. はじめに

我々は Web 上の人物の選択と理解を支援するために，Web 上の人名検索結果の要約と可視化研究を行っている[村上 09]．その中で，人物に位置情報，職業[上田 09]，NDC（図書館の分類記号）[村上 16]，NDLSH（国立国会図書館の件名標目）[下倉 17]等を付与してきた．本研究では，人物に日本図書館協会の基本件名標目表（BSH）第 4 版（BSH4）の件名標目を付与する．人物に BSH4 を付与することにより，精度の高い（ゴミの少ない）キーワードが付与でき，探索的検索等の応用が可能である．

本稿では，Web における人名検索結果から得られた Web ページに BSH4 を付与する方法を検討した．Web 上の 80 人物に対して，検索ランキング，文書内の位置，参照語，文書頻度の 4 種類を組み合わせた 405 パターンについて比較した実験結果を報告する．

2. BSH

BSH（Basic Subject Headings）は，日本図書館協会が提供する件名標目表であり，最新は第 4 版[日本図書館協会 99]である．件名標目は，目録を検索する手がかりとして提供されており，資料の主題をことばで表現したものである．BSH 第 4 版（BSH4）には，件名標目，参照語，説明付き参照，細目がある．また，各項目の中に最上位標目，上位標目，下位標目，直接参照，直接参照有り，関連標目等が含まれている．

3. 方法

3.1 手法

先行研究[佐藤 05]で使われた 20 の日本人氏名を用いて，Google Web APIs で 50 件の検索を行い，検索結果から同姓同名人物を手動で分離した 80 人分の Web ページ（HTML 文書）を利用した．人物毎の HTML 文書に対し，BSH4 の標目を付与する．

まず，BSH4 の件名標目のみを抜き出し，そこに直接参照があれば参照語を抽出する．

標目と参照語について，文字列が長い方がより詳細な意味を付与できると考えて，文字列の長いものから順に，以下の(a)と(b)で与えられるタグを除いた HTML 文書と照合してカウントする．一致した箇所は半角空白 1 つに置き換えて，次の標目または参照語を処理する．たとえば文書中の「人工知能」という文字列を処理する際に標目「人工知能」はカウントされるが標目「知能」はカウントされない．標目や参照語をカウントした後に重み付けを行い該当する標目のスコアを算出する．

組合せ条件として以下の 4 種類を用意した．

(a)Web ページの検索ランキングの利用：人物毎の上位 1, 3, 5, 10 件および全件の 5 パターン．

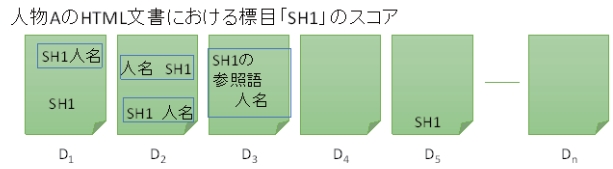
(b)HTML 文書内の位置の利用：タイトル，全文，検索語（人名）の前後の文字（前後 20, 40, 60, 80, 100, 150, 200）の 9 パターン

(c)参照語の利用 ：参照語を利用しない，標目の 0.5 倍の重みで利用，標目と同じ重みで利用の 3 パターン

(d)標目および参照語の文書頻度の利用：何もしない，文書頻度（df）/利用した全文書数（N）をかける，利用した全文書数（N）/文書頻度（df）をかけるの 3 パターン

これらを組み合わせると 5×9×3×3=405 パターンとなる．図 1 に標目のスコア計算例を示す．

最上位のスコアを持つ標目を該当人物に付与する．ない場合は「なし」とする．



パターン (a) 上位5件，(b) 人名の前後x文字，(c) 文書頻度/全文書数，(d) 参照語0.5倍
SH1: 3回出現；SH1の参照語：1回出現×0.5=0.5
(3+0.5)×3/5=3.5×0.6=2.1

パターン (a) 上位3件，(b) 全文，(c) 全文書数/文書頻度，(d) 参照語なし
SH1: 4回出現
4×3/2=4*1.5=6.0

図 1：スコア計算例

3.2 評価

人物に付与する最も適当な BSH4 の標目を著者らが 1 つ選定し，正解データとした．たとえば，元野球選手の江川卓氏は標目「野球」，関西学院大学教授の三浦麻子氏は標目「社会心理学」とした．80 人物全員に付与できた．

評価指標は以下のとおりとする．

$$\text{正解率} = \frac{\textit{自動的に付与されたBSHが正しい人物数}}{\textit{人物数}}$$

連絡先：下倉 雅行，大阪市立大学大学院創造都市研究科，大阪市住吉区杉本 3-3-138，masayuki@karekinada.com

$$\text{適合率} = \frac{\textit{自動的に付与されたBSHが正しい人物数}}{\textit{自動的にBSHが付与された人物数}}$$

$$\text{再現率} = \frac{\textit{自動的に付与されたBSHが正しい人物数}}{\textit{手動で付与された正しいBSHがある人物数}}$$

$$\text{F 値} = \frac{2 \times \textit{適合率} \times \textit{再現率}}{\textit{適合率} + \textit{再現率}}$$

$$\text{総合精度} = \frac{\textit{正解数}}{\textit{人物数}}$$

ただし，総合精度の正解数は手動で付与できなかった人物に対して自動で付与できない場合は正解とするが，今回の場合は手動で付与できない人物がいないため，正解率と同じとなる．

4. 方法

正解率においてどのパターンが最も良かったかを表 1 に示す．

全体（80 人物）について最も良かったパターンは，「上位 10 件，全文を利用し，参照語なし，df/N」であり，正解率は 27.5%（22/80）であった．

人物毎の文書数によって傾向が異なることが観察されたため，1, 2, 3 文書以上，11 文書以上，3 文書以上 10 文書以下に分けて調べた．1 文書しかない 35 人物の場合「全文（参照語は利用しない）」が最も良かった．3 文書以上の 33 人物の場合，「上位 10 件，人物名の前後 150 文字を利用し，参照語は標目の半分の重みで，df/N を利用する」が最も良かった．

表 1 正解率の高いパターン

文書数	上位件数	利用箇所	参照語	文書頻度	正解率
全体	10	全文	0	df/N	0.275 (22/80)
1	1	全文	0	1	0.286 (10/35)
2	3	前後 200	0	1	0.333 (4/12)
3 以上	10	前後 150	0.5	df/N	0.333 (11/33)
11 以上	5	前後 40	1	df/N	0.389 (7/18)
3～10	10	前後 80	0	df/N	0.267 (4/15)

表 2 に正解率の高いパターンの適合率，再現率，F 値，総合精度を示す．全体的に 30%前後という結果が得られた．

表 2 正解率の高いパターンの評価

文書数	適合率	再現率	F 値	総合精度
全体	0.278	0.275	0.277	0.275
1	0.294	0.286	0.290	0.286
2	0.333	0.25	0.286	0.250
3 以上	0.343	0.333	0.338	0.333
11 以上	0.389	0.389	0.389	0.389
3～10	0.286	0.267	0.276	0.267

比較のためにベースラインとして余弦を用いた方法を実装した．MeCab を用いて標目と参照語と文書から名詞を抽出して用語とした．全体で最も良かったパターンと条件を合わせるために，上位 10 件，全文，参照語なし，文書頻度を用いた．標目から抽出した用語の重みは出現頻度×df/N とした．その結果，余弦を用いた方法では正解が得られなかった．

以上より，全体としては，Web ページ全てよりも上位 10 件に抑え，HTML 文書の全文を利用し，参照語を利用せず，多くの文書に出現する語に重み付けすると良いことがわかった．

5. 関連研究

統制語を文書に付与する方法は機械学習を行うものと行わないものに大別され，本研究は機械学習を行わないものである．機械学習を行わない場合，余弦がベースラインの一つとなっている．余弦を利用しても正解が得られなかったことを考えると本研究では良い結果が出ていると言える．

NDLSH を人物に付与する研究[下倉 17]において，全体として良いパターンは「上位 10 件，人物名の前後 100 文字，同義語は標目の半分の重みで，文書頻度 df/N を利用する」であった． NDLSH における同義語と BSH4 における参照語とは基本的に同じである．すなわち，上位 10 件と文書頻度 df/N を利用する点は同じであり，利用箇所と参照語については異なった．結果の相違点については，BSH4 と NDLSH の収録語数の違いによる可能性がある．本研究で使用した標目の収録数は，BSH4 は 7847 件，NDLSH は 2016 年 9 月現在で 19521 件であった．今後詳細な検討が必要である．

6. おわりに

Web における人名検索結果から得られた Web ページに BSH4 を付与する方法を検討した．Web 上の 80 人物に対して検索ランキング，文書内の位置，参照語，文書頻度を組み合わせた 405 パターンについて比較実験を行った．結果として，上位 10 件の文書の全文を利用し，参照語を利用せず，標目の出現文書数を利用した全文書数で割ったものをかけたものを利用することが一番よいということがわかった．

今後の課題としては，文書数ごとに適したパターンのさらなる調査，上位標目や下位標目等の利用，最上位以外のデータの検証，実験データの追加，等があげられる．

参考文献

[村上 09] 村上 晴美，上田 洋：Web 人名検索結果の要約と可視化を目指して: 2009 年度人工知能学会全国大会（第 23 回）論文集 (2009)

[上田 09] 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)

[村上 16] 村上 晴美，浦 芳伸，片岡 祐輔: Web 上の人物への図書館の分類記号の付与と人物ディレクトリの開発，システム制御情報学会論文誌, Vol.29, No.2, pp.51-64 (2016)

[下倉 17] 下倉 雅行，村上 晴美：Web 上の人物への NDLSH の付与，人工知能学会全国大会（第 31 回）論文集 (2017)

[日本図書館協会 99] 日本図書館協会件名標目委員会編，基本件名標目表 第 4 版，日本図書館協会 (1999)

[佐藤 05] 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, pp.26-36 (2005)