1L3-1

Web 上の人物への NDLSH の付与

Assigning NDLSH Headings to People on the Web

下倉 雅行　　村上 晴美
Masayuki Shimokura　　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 National Diet Subject Headings (NDLSH) to the results of web people searches to help users select and understand people on the web. By assigning NDLSH 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) document frequency, and (d) synonyms. 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]等を付与してきた．本研究では，人物に国立国会図書館の件名標目である NDLSH を付与する．人物に NDLSH 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 NDLSH 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，同義語，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. NDLSH

NDLSH（National Diet Library Subject Headings）は，国立国会図書館が提供する件名標目表である．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．件名標目（以下標目），標目よみ，ID，同義語，上位語，下位語，関連語，注記，分類記号（NDLC），分類記号（NDC9），参照（LCSH），参照（BSH4），出典（BSH4），出典，編集履歴，作成日，最終更新日の 17 の項目で構成されている（http://id.ndl.go.jp/information/download/）．

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，NDLSH の標目を付与する．

まず，NDLSH の標目と同義語を抽出する．この時，標目からは半角英数字 2 文字以下，全角 1 文字のみ，ーー（ハイフン 2 つ）が含まれる語はあまり重要ではないあるいは照合がうまくいかないと考えて除去している．

標目と同義語について，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，次の標目または同義語を処理する．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．標目や同義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．

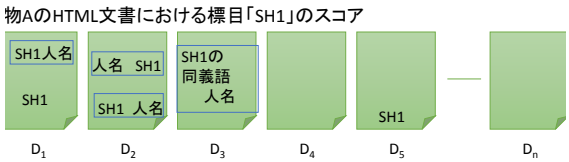
(b) HTML 文書内の位置の利用：タイトル，全文，検索語（人名）の前後の文字（前後 20, 40, 60, 80, 100, 150, 200）の 9 パターン

(c) 同義語の利用 ：同義語を利用しない，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) 標目および同義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．



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) 同義語0.5倍
SH1: 3回出現；SH1の同義語：1回出現×0.5=0.5
(3+0.5)×3/5=3.5×0.6=2.1

ーン (a) 上位3件，(b) 全文，(c) 全文書数/文書頻度，(d) 同義語なし
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 NDLSH 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．79 人物に付与できた．

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

評価指標は以下のとおりとする.

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

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

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

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

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

ただし, 総合精度における正解数には, 正しい NDLSH が存在しない人物に自動的に NDLSH を付与しない場合を含む.

4. 結果と考察

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, 人名の前後 100 文字(合計 200 文字), 同義語 0.5 倍, df/N」であり, 正解率は 26.3%(21/80)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(同義語は利用しない)」が最も良かった. 3 文書以上の 33 人物の場合全体とほぼ同じ(違いは同義語の倍率のみ)であり, 11 文書以上の 8 人物と同じ結果であった.

表 1 正解率の高いパターン

文書数	上位件数	利用箇所	同義語	文書頻度	正解率
全体	10	前後100	0.5	df/N	0.263 (21/80)
1	1	全文	0	1	0.286 (10/35)
2	3	前後200	0.5	1	0.333 (4/12)
3以上	10	前後100	1	df/N	0.364 (12/33)
11以上	10	前後100	1	df/N	0.500 (9/18)
3～10	5	前後60	0.5	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. 全体及び 1 文書の人物では適合率, 再現率ともに 30%弱であるが, 11 文書以上の人物では 50%となっている.

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 を用いて標目と同義語と文書から 2 文字以上の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, 前後 100 文字, 同義語, 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, 同義語から抽出した用語の重みは出現頻度の 0.5 倍×df/N とした.

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	0.276	0.266	0.271	0.263
1	0.286	0.294	0.290	0.286
2	0.333	0.333	0.333	0.333
3 以上	0.364	0.364	0.364	0.364
11 以上	0.500	0.500	0.500	0.500
3～10	0.182	0.267	0.216	0.267

表 3 余弦を用いた場合の評価

文書数	適合率	再現率	F 値	総合精度
全体	0.026	0.025	0.026	0.025

余弦を利用した場合に比べると正解率の高いパターンの性能が大幅に良いことがわかる.

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 文書の全文よりも人名の前後の文字列を利用し, 同義語を利用し, 多くの文書に出現する語に重み付けするとよいことがわかった. ただし, 1 文書しかない人物については全文から標目をカウントするだけが最も良かった. これらは, 余弦を用いた方法よりも大幅に性能が良かった.

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. 本研究で得た結果はベースラインを大幅に上回っている.

人物ディレクトリを開発するために NDC を Web 上の人物に付与する研究[村上 16]においては, 文書の利用箇所×手法の 2 条件で検討した. 利用箇所としてはタイトルが最も良く, 本研究とは結果が異なるが, これは研究目的とデータセットの違いによるものであると考える. 人物毎の文書数が 1, 2, 3 以上で良い手法が異なる可能性を示しており, 本研究に関連する.

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 NDLSH 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, 同義語, 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, 人名の前後 100 文字, 同義語を利用し, 文書頻度で重み付けする方法が良いことがわかった.

今後の課題としては, 標目の不要語処理の検討, 最上位以外のデータの検証, 実験データの追加, などがあげられる.

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)

[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)