D-019

NDLSH を用いた Web 上の人物探索システム
A System to Search for People on the Web Using NDLSH

白川 欣岳† 下倉 雅行‡ 村上 晴美†
Yoshitaka Shirakawa Masayuki Shimokura Harumi Murakami

1. はじめに

人物検索は Web 検索において重要な課題の一つである。Web 上の人物検索においては、同姓同名人物の混在などにより目的の人物にたどりつけないことや、人物を識別できないことがある。人物の検索や識別のために、人物を特徴付けるラベル付けが重要である。本研究では、国立国会図書館の件名標目である NDLSH をラベルとして人物に付与する。NDLSH を付与することにより、上位語、下位語、関連語を用いた探索的な検索が可能となる。そして、NDLSH を視覚的に探索しながら Web 上の人物を検索するプロトタイプシステムを試作した。以下、2 節では NDLSH について、3 節では Web 上の人物に NDLSH を付与する手法について、4 節ではプロトタイプシステムについて述べる。

2. NDLSH

NDLSH（National Diet Library Subject Headings、国立国会図書館件名標目表）は国立国会図書館が資料の主題検索のために作成した件名標目表であり、件名標目、標目読み、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)、分類記号(NDC10)、参照(LCSH)、参照(BSH4)、出典(BSH4)、出典、編集履歴、作成日、最終更新日の 18 項目で構成されている。

3. Web 上の人物への NDLSH の付与

3.1 先行研究

下倉らは Web 上の人物に NDLSH を付与する手法の検討を行い、最上位の結果で評価を行った[1]。本研究はこの研究をベースとするため、以下に概要を述べる。

3.1.1 データセット

佐藤らの研究[2]で使われた 20 の日本人氏名を用いて Google Web APIs でそれぞれ 50 件の検索結果を取得し、同姓同名の人物を手動で分離した 80 人分の Web ページ（HTML 文書）をデータセットとする。

3.1.2 付与手法

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、--（ハイフン 2 つ）が含まれる語は除去している。原則として文字列の長いものから順に HTML 文書と照合してカウントし、一致（カウント）した箇所は空白に置き換える。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。組み合わせ条件として以下の 4 種類を検討した。5×9×3×3=405 パターンとなる。

(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

3.1.3 最上位の評価

著者らが正解データを与えて最上位データと一致するか評価を行った。最も良かったパターンは「(a)上位 10 件、(b)人名の前後 100 文字（合計 200 文字）、(c)同義語 0.5 倍、(d)df/N」であり、正解率は 26.3%（21/80）であった。このパターンを先行研究の最上位最良パターンと呼ぶ。

3.2 複数件名の付与

人物の検索、識別のために件名をラベルとして付与する場合、最上位 1 件だけではなく複数件必要である。本研究では人物に対して複数の件名を付与する。

まず、先行研究の最上位最良パターンの採用を検討した。抽出された標目と人物との関連度を著者らが人手で 1-5 の 5 段階(5：非常に関連している；4：やや関連している；3：どちらとも言えない；2：あまり関連していない；1：全く関連していない)で上位 20 件まで評価した。抽出された標目を観察したところ（表 1 に上位 10 件の例を示す）最上位以外に良い標目や不要語が存在することがわかる。また、関連度が 1 で複数回出現している標目の多くは同義語の利用によるものであることがわかった。

表 1　最上位最良パターン

田中克己 00	菱沼聖子 00	五斗進 00
データベース	動物	化学
大学	ハム	大学
情報処理	イトウ	**バイオインフォマティクス**
学術団体	医師	教育
工学	**獣医学**	薬学
オブジェ	公衆衛生	生命科学
タラ	ヒスチジン	データベース
テン	仮名	イギリス人
編集	時間・空間	日本人
LAN	スキー	科学

そこで、最上位最良パターンから同義語の利用をやめるもの、すなわち「(a)上位 10 件、(b)前後 100 文字、(c)同義語なし、(d)df/N」を同義語なしパターンと呼び、採用を検討する。表 2 に同義語なしパターンの上位 10 件を示す。全体的に不要語が減っていること、「イトウ」「イギリス人」「日本人」等の同義語利用による不要語がなくなっている。

†大阪公立大学 Osaka Metropolitan University
‡大阪市立大学 Osaka City University

Copyright © 2022 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

表 2　同義語なしパターン

田中克己 00	菱沼聖子 00	五斗進 00
データベース	動物	化学
大学	**獣医学**	大学
情報処理	公衆衛生	教育
工学	スキー	生命科学
オブジェ	会社	薬学
技術	体質	**バイオインフォマティクス**
編集	大学	データベース
研究機関	学生	生物
爆発	小説	生命
学生	博士	遺伝子

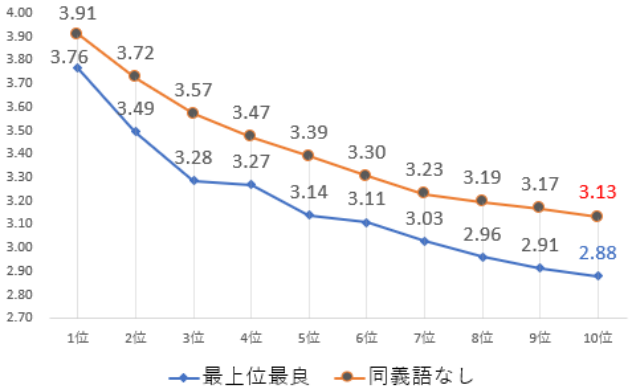

図 1　累積関連度

2 パターンの評価を累積関連度（該当順位までの関連度の平均）により行った（図 1）。最上位から 10 位まで、同義語なしパターンの方が良いことがわかる。

今回は、累積関連度が 3 を超えている上位 5 件の NDC を人物に付与した関連研究[3]を参考に、同義語なしパターンの上位 10 件の NDLSH を付与することにした。

件名を複数付与する場合は、先行研究の最上位最良パターンよりも、同義語以外の条件が同じで同義語を利用しないパターンの方が良かった。ただし、正解データの順位が下がることもあり、さらなる精度の改善が必要である。

4. プロトタイプシステム

件名標目を視覚的に探索するシステム Subject World on the Web[4]をベースに、NDLSH を視覚的に探索しながら Web 上の人物を検索するプロトタイプシステム People on the Web を試作した。標目の上位語、下位語、関連語を探索しながら人物検索が行える。データセットの 80 人物に同義語なしパターンで上位 10 件の標目を付与した。

4.1 機能と使用方法

検索窓でキーワードを入力するとキーワードを含む NDLSH の標目一覧が表示され、選択した標目がウインドウに表示される。標目をダブルクリックするとコンテキストメニューが表示される。そこで「NDLSH」を選択するとその標目の上位語、下位語、関連語の標目が表示され、「PEOPLE」を選択するとその標目が付与されている人物が表示される。表示された人物を選択すると右側の領域でその人物に付与された標目を確認できる。また、人物をダブルクリックしてコンテキストメニューから「SUBJECT」を選択すると、人物に付与された NDLSH の標目を表示できる。標目や人物のオブジェクトは、移動、拡大、縮小等が可能である。

4.2 使用例

図 2 に使用例を示す。「情報検索」という標目で検索を行い NDLSH を探索的に検索した。「データベース」や「情報処理」という標目は「情報検索」の関連語や上位語であり、「データベース」と「情報処理」が付与されている人物を表示させている。また、「データベース」が付与されている人物の一人である五斗進氏を選択すると右側の領域に付与された標目が表示され、「バイオインフォマティクス」等の詳細な標目が確認できる。さらに、五斗進氏に付与された別の標目を表示し、「遺伝子」が付与された野村紀子氏を表示させている。

このように、入力したキーワードから関連のある件名を探索することにより人物を検索したり、人物に関連のある件名を通して別の人物を検索することができる。

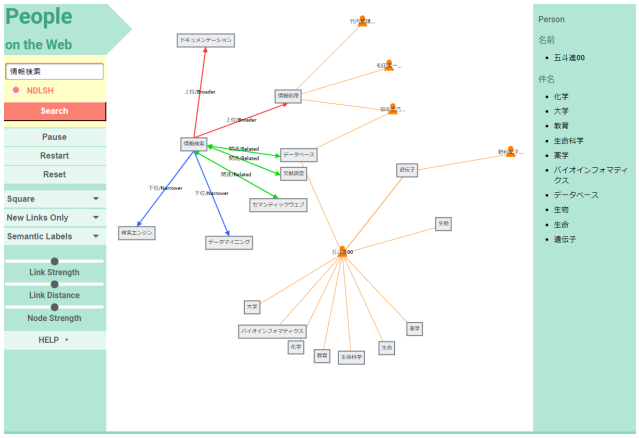

図 2　プロトタイプシステム

5. おわりに

人物の検索や識別のために人物に NDLSH を複数付与する手法を検討し、NDLSH を視覚的に探索しながら Web 上の人物を検索するプロトタイプシステムを試作した。今後の課題としては、付与手法の精度の改善、プロトタイプシステムの機能やデータの追加等があげられる。

謝辞

本研究は JSPS 科研費 19K12718 の助成を受けたものです。

参考文献

[1] 下倉雅行, 村上晴美, “Web 上の人物への NDLSH の付与”, 2017 年度人工知能学会全国大会(第 31 回)論文集 (2017).
[2] 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).
[3] 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 (2016).
[4] 村上晴美, “Subject World on the Web”, 第 66 回日本図書館情報学会研究大会発表論文集, pp. 123-124 (2018).

Copyright © 2022 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.