

Web を対象としたプロフィール情報抽出の基礎的検討

大前 信弘[†] 吉谷 仁志[‡] 黄瀬 浩一[‡] 松本 啓之亮[‡]大阪府立大学工学部情報工学科[†]大阪府立大学大学院工学研究科情報工学分野[‡]

e-mail : {ohmae, yoshitani}@ss.cs.osakafu-u.ac.jp, {kise, matsu}@cs.osakafu-u.ac.jp

1 はじめに

近年のインターネットの普及により, Web 上には大量の電子文書が存在し, 大規模なデータベースとなっている. Web の規模が大きくなるにつれて, 目的とする情報を自動的に収集し, まとめて欲しいという要求が高まってきている. とりわけ, 人物に関する情報(プロフィール情報)に関するそれは高い. しかし, 従来の情報抽出の研究では, 新聞記事などが研究対象の中心であり, Web ページを対象とした研究はまだ十分なされていない.

本稿ではプロフィール情報を対象とし, Web ページから情報抽出を行うための基礎的技術である「フィルタリング」と「切り出し」の処理について新しい手法を提案する. 「フィルタリング」とは, Web ページにプロフィール情報が含まれているかどうかを判別する処理であり, 「切り出し」とは, Web ページのどの部分にプロフィール情報が含まれているかどうかを判別する処理である. 本手法の特徴は, 両者にサポートベクトルマシンを用いる点にある. Web ページ 400 ページを対象として実験を行った結果, F 値 0.718 を得た.

2 プロフィール情報の抽出方法

情報抽出の手順を図 1 に示す. 本稿では, 「フィルタリング」と「切り出し」の 2 つの処理を提案する. 残りの 2 つの基礎技術である「固有表現抽出」と「情報統合」の処理については, 今後の課題とする. 以下で「フィルタリング」, 「切り出し」の手法を説明する.

2.1 フィルタリング

フィルタリングの手順を以下で説明する.

step1 HTML タグの除去

HTML タグにはプロフィール情報を表す情報は含まれていないので取り除く.

step2 形態素解析

形態素解析を行い文を単語列に分割する.

step3 ベクトル化

Web ページをベクトルで表す. 方法は, 以下の通りである. まず, ページに出てきた単語にベクトルの 1 つの要素を割り当てる. 次に, 局所的重み付け TF と大域重み付け IDF を用いてベクトルの要素に値を与える. 最後に, 各重み付けをされたベクトルを正規化する. TF, IDF, 正規化には以下の式を用いる.

Fundamental Investigation of Profile Information Extraction from the World Wide Web

N. Ohmae[†], H. Yoshitani[‡], K. Kise[‡] and K. Matsumoto[‡]

[†]College of Engineering, Osaka Prefecture University

[‡]Graduate School of Engineering, Osaka Prefecture University

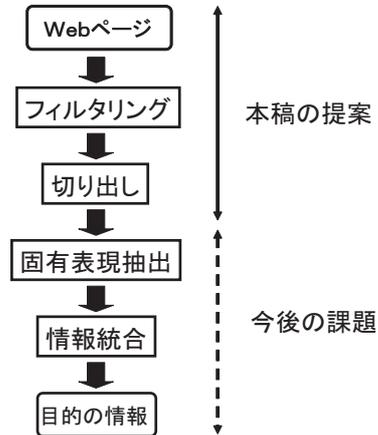


図 1: Web からの情報抽出の手順

- TF

$$l_{ij} = \log(1 + f_{ij})$$

- IDF

$$g_i = \log \frac{n}{n_i}$$

- 正規化

$$v'_j = \frac{v_j}{\|v_j\|}$$

ここで, f_{ij} は単語 w_i の Web ページ d_j における出現頻度, n は Web ページの総数, n_i は単語 w_i を含む Web ページ数, v_j は局所重み付けと大域重み付けを行った後のベクトル $(l_{1j}g_1, \dots, l_{mj}g_m)$, $\|v_j\|$ は, ベクトル v_j のノルムである.

step4 は, 学習の場合と判別の場合で異なる.

step4 学習

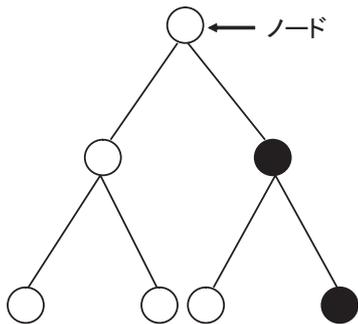
プロフィール情報の有無とベクトルの関係をサポートベクトルマシン (SVM) [1] に学習させる. このとき, ベクトル化された Web ページが SVM の素性ベクトルとなり, プロフィール情報を持つかどうかラベルとなる.

step4 判別

ベクトル化した Web ページを素性ベクトルとして SVM に与える. そして, プロフィール情報を持つかどうかを判別させる.

2.2 切り出し

切り出しの手法では, Web ページを図 2 のような HTML タグによる木構造として解析する. そして, プロフィール情報の切り出しを, プロフィール情報を含んでいるノードを選



○プロフィール情報なし ●プロフィール情報あり

図 2: タグによる木構造

表 1: 実験結果

実験	再現率 R	精度 P	F 値
フィルタリング	0.895	0.894	0.895
切り出し	0.732	0.732	0.732
フィルタリング後の切り出し	0.700	0.737	0.718

択することとみなす。ノードにプロフィール情報が含まれているかどうかを判別するには、そのノードを根とする部分木に含まれる単語を用いる。ページ内のすべてのノードを判別し、結果を次のように扱う。「フィルタリング」と同様に、そのベクトルを正規化したものを素性ベクトルとして SVM に学習させる。そして、ノードがプロフィール情報を持つかどうかを判別する際も、ノードの持つ単語から作られるベクトルを素性ベクトルとして SVM に与えて判別する。IDF と正規化の式には 2.1 と同じ式を用いる。一方、TF には、 f_{ij} を単語 w_i のノード n_j における出現頻度に置き換えたものを用いる。

3 実験

本研究での提案手法である「フィルタリング」、「切り出し」の有効性を検討するために以下の実験を行った

3.1 フィルタリング

Web ページにフィルタリングを施し、プロフィール情報が含まれているページを抜き出す実験を行った。形態素解析器には茶筌 [2] を、SVM には TinySVM [3] を用いた。SVM の仕様は、2 次の多項式カーネルを用い、ソフトマージンを 1 とした。この実験には、Google によって「名前 年齢 自己紹介」というキーワードで検索した Web ページを用いた。内訳は、プロフィール情報が含まれている Web ページ上位 200 ページ、含まれていない Web ページ上位 200 ページである。これらの Web ページを用いて、10 分割交差検定を行った。実験の評価には、再現率 $R = |C|/|A|$ 、精度

$P = |C|/|B|$ 、再現率 R と精度 P で表される F 値 $F = \frac{2RP}{R+P}$ を用いた。ここで、 $|A|$ はプロフィール情報を含むページ数、 $|B|$ は結果として得られたプロフィール情報を含むページ数、 $|C|$ は A 内の正解数である。実験結果を表 1 に示す。

この実験で誤って「プロフィール情報が含まれていない」と判別されたページには、プロフィール情報が「1981・12・4 生まれ いて座 A 型 高知出身」などのように省略形で書かれていて、プロフィール情報を表す手がかりが少ないという特徴が見られた。逆に、誤って「含まれている」と判別されたページには、「氏名」、「年齢」などの単語が多数含まれているという特徴が見られた。

3.2 切り出し

Web ページからプロフィール部分を切り出す実験を行った。SVM には前述の TinySVM を用いた。ここでは、RBF カーネルを用い、ソフトマージンを 1 とした。この実験には、プロフィール情報が含まれている Web ページのみの 200 ページ用いて、10 分割交差検定を行った。実験の評価には、4.1 で $|A|$ をプロフィール情報を含むノード数、 $|B|$ を結果として得られたプロフィール情報を含むノード数、 $|C|$ を A 内の正解数に置き換えた再現率、精度、 F 値を用いた。実験結果を表 1 に示す。

この実験での誤りは、プロフィール情報を持たない親ノードとプロフィール情報を持つ子ノードとの間に、含まれている単語の差が少ないときに多く見られた。

3.3 フィルタリング後の切り出し

フィルタリングの結果を対象に、切り出しを適用する実験を行った。実験結果を表 1 に示す。

この結果は、「切り出し」を単独で行った結果とほぼ同じ F 値であった。このことから、誤りを含んだ「フィルタリング」結果に対して、「切り出し」では誤りにあまり影響されることがなくノードのプロフィール情報の有無を判別出来ていることがわかる。

4 おわりに

本研究では、Web ページから情報抽出するための基礎技術である「フィルタリング」、「切り出し」の手法を提案した。今後、残りの基礎技術である「固有表現抽出」、「情報統合」について研究を行いたい。

謝辞 本研究は日本学術振興会科学研究費補助金 (C)(14580453) の補助による。

参考文献

- [1] 前田英作：痛快！サポートベクトルマシン：情報処理学会誌，Vol.42，No7，pp. 676–683，2001.
- [2] URL : <http://chasen.aist-nara.ac.jp/>
- [3] URL : <http://cl.aist-nara.ac.jp/~taku-ku/software/TinySVM/>