2G-5

Webを対象としたプロフィール情報の項目化と統合

吉谷 仁志† 黄瀬 浩一† 松本 啓之亮†
大阪府立大学大学院工学研究科情報工学分野†
e-mail : yoshitani@ss.cs.osakafu-u.ac.jp, {kise, matsu}@cs.osakafu-u.ac.jp

1 はじめに

Web ページの増加に伴い，目的の情報をいかに効率よく収集できるかに大きな注目が集まっている．従来の情報検索技術の多くは結果をページ単位で返すため，それらを 1 つ 1 つ読む手間が発生する．これを解決する手法として，短い文章形式にまとめる自動要約や表形式にまとめる情報抽出が提案されている．

自動要約は結果が文章形式であるため，表現の自由度は高いが一覧性が低いという問題点がある．一方，情報抽出は結果が表形式なので一覧性は高いが,「出身地－大阪」などのような属性－属性値の対応関係が必要であり，自由度は低い．このため,「通算成績」や「デビュー時期」のような特定のジャンルに依存する属性をあらかじめ指定できない対象について情報抽出を適用するのは困難である．

本研究では，属性－属性値の関係が必要なく，文章形式より一覧性の高い箇条書きの形式で目的の情報を集約する手法を提案する．箇条書きの抽出に関しては，手順に関するものを対象とした研究 [1] があるが，これはあらかじめ箇条書きで書かれた部分のみを対象としている．本研究では，対象文書の形式によらず情報を項目化して抽出する．対象の情報としては，人物に関する情報に注目が集まっていることからプロフィール情報 [2] を考える．

2 プロフィール情報の項目化と統合

提案手法では，目的の人物名を入力として，その人物に関するプロフィール情報を箇条書きの形式で提示する．その手順は Web ページの選別，項目抽出，項目の統合の 3 つからなる．以下で各手順の詳細な処理手順について述べる．

2.1 Web ページの選別

提案手法ではまず，対象人物名を検索質問として検索エンジンから Web ページを収集する．この際，得られたページには対象人物のプロフィール情報が存在するページとそうでないページが存在する．Web ページの選別では，これらのページから前者のみを選別する．

まず，図 1 に示すように，Web ページを木構造で表現する．次に，解析した木の各ノードに人名が含まれているかどうかを判定し，人名が含まれているノードを候補ノードとする．最後に，候補ノードを幅優先探索順にソートし，以下の基準で選別する．対象人物名のみが含まれている場合は，対象人物の情報があるページと判断する．対象人物とは異なる名前のみが含まれている場合は，他人のページと判断する．双方の名前が存在する場合は，次の候補ノードに判定を委ねる．なお，全ての候補ノードに双方の名前が存在する場合や候補ノードが存在しない場合はプロフィール情報なしと判断する．

Itemization and Integration of Profile Information from Web Pages.
Hitoshi Yoshitani†, Koichi Kise† and Keinosuke Matsumoto†
†Graduate School of Engineering, Osaka Prefecture Univ.

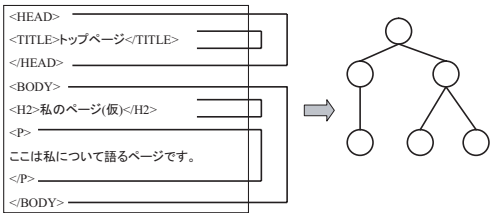

図 1: HTML の木構造表現

2.2 項目抽出

項目抽出では，選別の結果得られたページに対して，プロフィール情報に関する部分を項目化して取り出す．まず，文章形式の部分に対しては文単位で区切り，表形式のものに対してはセル単位で分割することで項目化を行う．ただし，セル内のテキストが属性表現のパターン (あらかじめ人手で作成) に合致する場合は，次のセルのテキストと結合する．以下では項目化された各テキストを項目候補とよぶ．

次に，項目候補の中からプロフィール情報を表すものを選択する．これには，各項目をベクトル化し，それをサポートベクトルマシン (SVM) とよばれる機械学習器に与えることで選択基準の学習及び識別をする．今，j 番目の項目をベクトル化したものを $\boldsymbol{v_j} = (c_{1j}, c_{2j}, \cdots, c_{mj})$ (m は単語の異なり数) とする．このとき，c_{ij} は情報検索で用いられる logTF および IDF 重みに基づく値を用いる．

2.3 項目の統合

項目の統合では，項目抽出の結果得られた項目のうち冗長な項目を 1 つにまとめる．まず，項目の表記ゆれをあらかじめ用意した辞書により統一する．次に，前節で述べた $\boldsymbol{v_j}$ を用いて，$\boldsymbol{v_p}$ と $\boldsymbol{v_q}$ の類似度 $sim(p,q) = \boldsymbol{v_p}^T \cdot \boldsymbol{v_q}$ を計算し，これがあらかじめ定めた閾値 th_i 以上になった組み合わせを統合候補の集合 N の要素とする．ただし，同一文書内の項目同士の組み合わせは除く．

統合の対象となる文書は以下に示す手順で選択する．最初に，N の中で文書 d_k と文書 d_l に関するものを選び，その類似度の総和を d_k と d_l の関連度と定義する．これを全ての文書の組み合わせについて計算し，その平均 avg と分散 sd を求める．次に，関連度が閾値 $th_d = avg + \alpha \cdot sd(\alpha$ は定数) 以上の組み合わせを用いて図 2 に示すような無向グラフを作成し，最大の連結部分グラフに含まれる文書を統合の対象とする．ここで，ノード内の番号は文書番号を表す．図 2 の場合は，左側のグラフに含まれる 1,2,3,4,10 を対象とし，それ以外の文書を削除する．

最後に，N のうち統合の対象に含まれるものに対して，項目を形態素解析し，記号を除いて完全に一致するもの及び一方が他方を包含しているものを統合する．統合すると判断された項目のうち形態素数が異なるものは，形態素数の多い方に統合する.「趣味　パソコン、読書、車」と「趣味　パソコン、映画鑑賞」のように一致しない形態素が存在するものについては，異なる情報と判断して統合しない．最後に，統

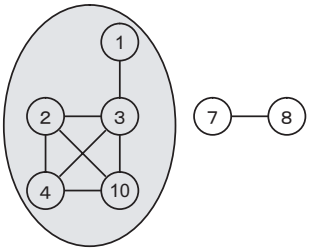

図 2: 統合する文書の選択

合された項目と，統合されなかった項目をあわせて出力し，最終結果とする．

3 実験

提案手法の有効性を検討するため，Web ページを対象とした項目化の実験を行った．まず，Web ページの選別実験をし，その結果を用いて項目抽出及び統合の実験をした．

3.1 Web ページの選別実験

提案手法による Web ページの選別実験を行った．対象文書は,「イチロー」,「小泉純一郎」,「宇多田ヒカル」,「所ジョージ」のそれぞれに「プロフィール」という検索語を付与し，Google で検索した際の各上位 20 件である．手法の有効性を評価する尺度としては，再現率 $R = |C|/|A|$，精度 $P = |C|/|B|$，F 値 $F = 2RP/(R+P)$ を用いた．ここで，$|A|$ は対象人物のプロフィール情報が含まれるページ数，$|B|$ は対象人物のページであると判定されたページ数，$|C|$ は $|B|$ 内の正解数である．

実験結果は平均で $R = 1$，$P = 0.61$，$F = 0.74$ となった．これより，選別によって目的の情報を失わずにある程度不要な Web ページを除去することに成功しているといえる．

誤りの傾向としては，Yahoo!カテゴリのように，プロフィール情報が含まれないページであるが，タイトル等に対象人物名が含まれていた場合が 38.4%，ハンドルネームなどが対象人物名と酷似していたものが 38.4%，人名の認定誤りによるものが 15.4%であった．

3.2 項目抽出の実験

選別された Web ページに対し，項目抽出の実験を行った．学習用データとして，Google で「プロフィール」,「自己紹介」を検索質問とした際に得られる Web ページからプロフィール情報を含むもの上位 50 件ずつを収集し，正解ラベルを与えた．ただし，Web ページの選別実験時に対象文書となったものは除いた．SVM の学習器として SVM^{light} [3] を用い，2 次の多項式 Kernel を使用した．結果は前節で述べた R，P，F で評価した．ただし，ここでの $|A|$ はプロフィール情報が含まれる項目候補数，$|B|$ はプロフィール情報が含まれると判断された項目候補数，$|C|$ は $|B|$ 内の正解数である．

実験結果は平均で $R = 0.71$，$P = 0.55$，$F = 0.61$ となった．誤りの傾向としては，メニューなどのプロフィール本体とは関係のない部分が全体の約 85%を占めた．そのため，プロフィール情報が書かれている部分のみを切り出した上で項目抽出をする必要があると考えられる．

表 1: 統合の実験結果

	対象人物の情報 (統合前→統合後)	その他の情報 (統合前→統合後)
異なり数	150 → 144	339 → 108
総数	236 → 179	366 → 114

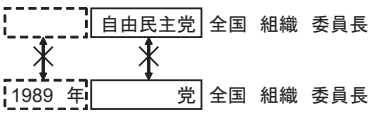

図 3: 統合できなかった例

3.3 統合の実験

抽出された項目に対して統合の実験をした．この際，$th_i = 0.5$，$\alpha = -0.5$ とした．統合の前後で対象人物のプロフィール情報を含む項目の異なり数と総数，それ以外の項目の異なり数と総数がどのように変化したかを調べた．この評価において，対象人物に関する項目の異なり数が統合の前後で変化せず，総数と一致した場合が理想的な統合といえる．また，その他の項目の異なり数，総数が少なければ少ないほど誤りを減らせているといえる．

結果を表 1 に示す．対象人物に関する項目の異なり数は統合の前後でほとんど変化しておらず，総数も異なり数に近くなっているので，有効な統合であると考えられる．また，その他の項目は異なり数，総数共に大幅に削減できているため不要な情報を効率よく削除できているといえる．

統合できなかった例としては，図 3 のように項目の情報が相補的になっている場合と,「趣味　食べること」と「趣味　読書，インテリア集め」などのように共通する情報がなく，統合ができなかった場合の 2 つが主であった．このため，項目を検索質問として再検索を行い，その結果から統合すべきかどうかを判断する必要がある．

4 おわりに

本稿では，Web ページから特定人物のプロフィール情報をまとめて提示する手法として，箇条書きによる項目化と統合の手法を提案した．本手法の特徴は，情報抽出のように属性を指定することなく，特定の事物に関する情報を項目形式でまとめる点にある．

今後の課題としては，項目抽出の精度を向上させることがあげられる．また，再検索の手法についても考える必要がある．最終的には対象人物名を入力するだけで，その人物のプロフィール情報が項目形式で提示されるようなシステムの構築を目指したい．

参考文献

[1] 武智峰樹，徳永健伸，松本裕治，田中穂積：“手順の説明を含む箇条書きを抽出するための手がかり分析”, 情処研報, NL-152-2, pp. 7-14, 2002.

[2] 吉谷仁志, 黄瀬浩一，松本啓之亮：“サポートベクトルマシンを用いた新聞記事からのプロフィール情報抽出”, 電学論 C, Vol.124, No.11, pp.2260-2266, 2004.

[3] URL：http://svmlight.joachims.org/