単語の頻度と位置に基づくプロフィール情報の抽出

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

1 はじめに

近年，インターネットの普及に伴って数多くの電子文書が存在するようになった．これら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)に関するそれは高く，これらを対象とした研究事例が報告されている [1]．しかし，従来のプロフィール情報の抽出は対象文書に特有な構造を利用するものもあり．一般性が十分達成されているとは言えない．

本研究では単語の頻度や位置といった対象文書に依存しない情報を用いて抽出を行う手法を提案する．また，特定の属性に依存しない情報統合を行う．

2 情報抽出

情報抽出は，電子文書から目的とする情報を抽出するために提案された手法である．この情報抽出は大きく分けて「個別属性の抽出」と「情報統合」の2つの処理からなる．なお，ここで言う「個別属性の抽出」のことを「情報抽出」と呼ぶ場合があるが，本稿ではこれに情報統合を加えたものを情報抽出と呼ぶことにする．

個別属性の抽出とは，文書中から人名や商品名といった特定の属性に対応する属性値を取り出す作業である．具体的には，対象文書にパターンマッチング(特定の語の存在から属性値となる語句を決定する処理)を施す．ここで，属性値を特定するための語を抽出パターンと呼ぶ．これを決定する方法としては人手によるものと機械学習によるものが提案されている．一般に大量の文書に対しては機械学習で高い精度が得られる [2]．

しかし，個別属性の情報は粒度が低いものであるため，それら単体の情報だけで有用な情報となることは少ない．そこで，これらの情報を整理してまとめることで情報の粒度を高め，有用な情報とすることを目的として情報統合が提案されている [3]．このような情報の整理・統合作業は有用な情報を作成する上で重要なものであると考えられる．

3 頻度と位置に基づく情報抽出

3.1 提案手法の方針

従来の情報統合手法では，住所や電話番号など特定の属性が利用できるものを処理対象とするものが多かった．これは住所や電話番号が一致すれば同じものに対する情報であると認定できるのに対し，それ以外の属性(出身地や最終学歴など)が一致しても同じものであるかどうかが分からないためである．しかし，それらが複数合わさることにより同じものに対する情報である可能性を高めることができる．例えば「1964年に大阪で生まれ，大阪府大を卒業した人」などがそうである．本研究ではこの点に着目することで，より一般性のある情報統合を提案する．

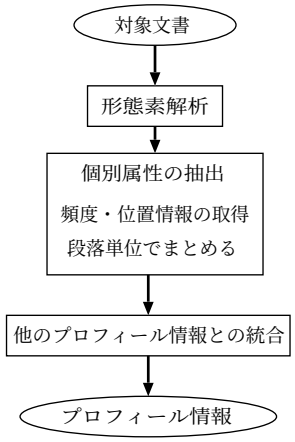

図 1: 提案手法の流れ

3.2 提案手法の流れ

以下で，図1に示す提案手法について説明する．初めに，対象文書に対して形態素解析を行い,「改行の後に字下げが行われている部分」を段落の切れ目として段落ごとの単語情報を得る．次に，得られた単語情報に対して人手により作成した抽出パターンを適用し，氏名，生年，出身地，学歴などの履歴の計4属性に対する属性値を抽出する．この際，既に得られた属性値情報と照合して，文字列が完全に一致するものを同じ属性値と判断することで属性値ごとの頻度を記録する．また，それと同時に各属性値の出現位置も記録する．この作業を段落の終りまで順に行う．

1つの段落に対する属性値の情報が得られたら，それらを同じ人物に対する情報とみなして属性値集合を作る．これは段落の区切りは比較的容易に分かることと，同じ段落内に複数人のプロフィール情報が書かれている可能性が低いからである．しかしこの方法では図2 のように，氏名情報のない属性値集合ができる場合がある．プロフィール情報では氏名の情報が重要であるため，氏名情報の欠落は極力避けたい．そこでこのような場合は，前の段落の中で最も頻度が高く，なおかつ最も前に位置している氏名情報をその段落の氏名とする．前の段落にも氏名情報がない場合は氏名情報のない段落としてまとめる．

段落ごとにまとめられた属性値集合ができると，それらを互いに比較することで情報の統合を行う．まず各属性の文

Extraction of Profile Information Based on Frequency and Location of Terms
Hitoshi Yoshitani†, Koichi Kise‡ and Keinosuke Matsumoto‡
†College of Engineering, Osaka Prefecture University
‡Graduate School of Engineering, Osaka Prefecture University

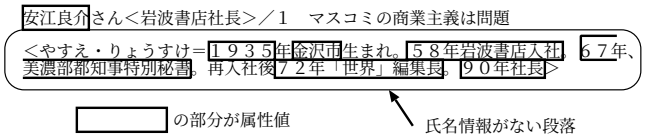

図 2: 氏名のない段落に対する処理

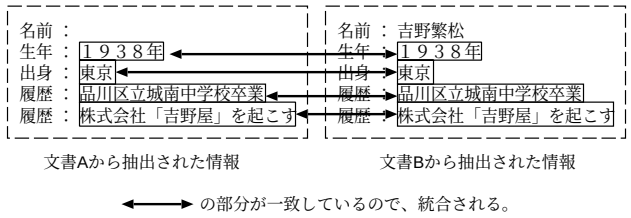

図 3: 情報の統合例

字列を比較し，完全に一致したものがある場合は属性ごとに定めた値を評価値に加える．属性ごとに定めた値については氏名の場合大きな値，その他の属性の場合は小さな値となるように設定する．このようにして各集合同士の評価値を計算し，評価値が一定値以上になれば同じ人物に対する情報として情報の統合を行う．このような手法を用いることで，どのような属性でも同一性の判定に用いることができる．

属性値集合の統合手順としては，通常は属性値の重複を取り除きながら互いの情報を統合する．各属性の属性値間で矛盾している項目がある場合は，それらの中で最も頻度が高く，なおかつ段落中の出現位置が最も前である属性値を統合後の属性値とする．これは,「頻度の高い情報ほど信憑性が高い」ということと「より前にある情報の方が正しいプロフィール情報である可能性が高い」という前提に基づいている．このような手法を用いることで，図3のように氏名属性が抽出できなかった情報を他の情報と統合することも可能になる．最後に，氏名のないプロフィール情報を削除して最終結果を出力する．

4　実験

本研究の有効性を検証するために，新聞記事を対象としたプロフィール情報の抽出実験を行った．まず，毎日新聞94年版1月分の記事より抽出パターンを作成し，次にそのパターンを用いて個別属性を抽出し統合するシステムを作成した．そして，作成したシステムを表1に示す条件で実行し，プロフィール情報を得た．

表 1: 実験条件

対象文書	毎日新聞94年版(2月分)
対象記事の選別	なし
記事数	約7400
抽出パターン数	21

表 2: 実験結果

抽出されたプロフィール数	56
精度	60.7%(34/56)
氏名の認定誤り	11
属性値の再現率	53.8%(92/171)
属性値の精度	90.2%(92/102)

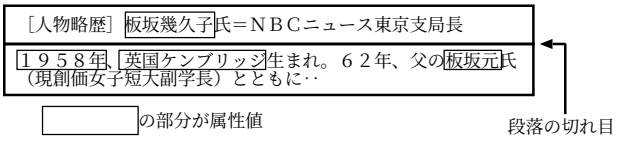

図 4: 誤った氏名が認定されている例

これを人手で作成した正解データと照合し，氏名属性の属性値が一致しているものを正解と判断した．また，正解とされたプロフィール情報に対し，どれだけ属性値が正しく抽出されているかを確かめるため，属性値の再現率と精度を計算した．なお，属性値は正解データのものと文字列が完全に一致したものを正解と判断する．このような判定基準に基づき，調べた抽出結果を表2に示す．

この結果より，本手法のように比較的簡単な特徴しか用いていなくてもある程度の結果は得られることが分かる．誤りと判断されたプロフィール情報の多くは図4のように，本来の氏名とは違ったものが氏名の属性値として認定されていたものであった．このことから，氏名の属性値については認定基準を見直す必要がある．

5　おわりに

本研究では，個別属性の抽出で得られた属性値情報を頻度と位置の情報に基づいて統合する手法を提案した．本手法の特徴は属性の種類に応じた個別処理によらず統合を実現している点にある．今後，機械学習を用いて個別属性の抽出及び情報統合の精度を向上させることで，システムの改善を行いたい．

参考文献

[1] 西野文人, 落谷亮. 新聞記事からの人物・企業情報の抽出. 情処研報,NL-127-17, pp. 125–132, 1998.

[2] 磯崎秀樹, 賀沢秀人. SVMに基づく固有表現抽出の高速化. 情処研報,NL-149-1, pp. 1–7, 2002.

[3] 佐藤理史, 佐藤円. 情報の自動編集とWITプロジェクト. 日本図書館情報学会研究委員会（編）, 電子図書館–デジタル情報の流通と図書館の未来, pp. 131–149. 勉誠出版, 2001.