1U-03

ファジィ積分に基づいたWeb文書の人物特定システム*

内海 慎久† 鈴木 英之進‡

utsumi@mtl.t.u-tokyo.ac.jp suzuki@slab.dnj.ynu.ac.jp

† 東京大学大学院工学系研究科 ‡ 横浜国立大学工学部電子情報工学科

1 はじめに

近年のネットワークの普及に伴い，WWW (World Wide Web) 上の情報は多分野にわたって急速に規模を拡大している．このように情報の氾濫している現在，検索エンジンはユーザーが必要な情報を得るために有用な手法と言える．

検索エンジンを利用する際に検索語に人物名を指定すれば，我々はその人物に関するWeb文書へのリンクを得ることができる．しかしリンクのなかには'Not Found'などの無効な文書や，地名のように検索語が人物名ではない文書 (非人物文書) へのリンクが含まれる．また各リンクは人物ごとに分かれて表示されないため，求める人物の情報を得るためには膨大な時間がかかってしまう．

この問題に対し，同一人物ごとにリンクをまとめて出力すれば，各人物に関する情報を効率的に得ることができる．そこで本稿では，検索エンジンで人物名を検索した場合に出力されるWeb文書群を人物ごとに分けるシステムを提案し，その有効性を評価する．10,000個のWeb文書に対して実験を行なったところ，平均93.6%でWeb文書を人物ごとに分け，システムの有効性を示した．

2 人物特定システムの構成

図1に人物特定システムの構成を示す．入力は人物の姓を検索語として既存の検索エンジンより出力されたWeb文書群である．システムは始めに各Web文書の姓が人物を示すかを判定し，その姓のファーストネームごとに文書を分類して初期クラスを生成する．ここでファーストネームを取得できなかった文書は一つのクラス (名無し) として表す．

次に各Web文書中の単語N個に重み付けをし，重みを要素とするベクトルとして表す．ここでNは文書により異なり，文書に出現する姓の周辺100語と文頭文末20語から得る．次に，初期クラス内の各2文書間の類似性から初期クラスを分割する．生成クラスにファーストネームを取得できなかった全文書を統合し，さらにクラス別に文書の類似性を測り，それをもって最終出力とする．

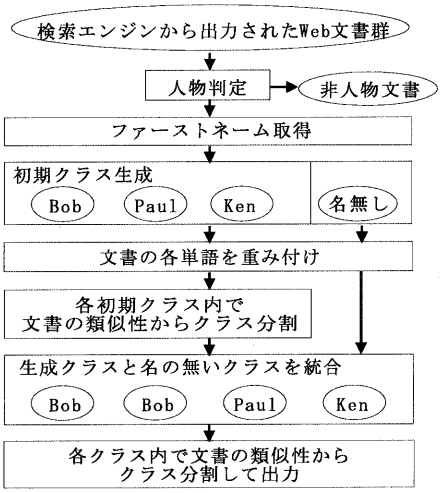

図 1: システムの構成

3 単語への重み付け

3.1 既存の重み付け手法

単語への重み付けは，人物を特定するために重要な作業である．代表的な手法には，頻度と文書単位の珍しさで重みを付けるTF-IDF法，関連単語数と排他性で重みを付ける関連単語群判定 [1] などがある．

WWWには名簿のような文書，図が多い文書，文字が少ない文書などが多く存在する．よって一般的単語がWeb文書に含まれるとは限らない．このため，TF-IDF法では一般的単語を大きく重み付けてしまう問題点がある．対して関連単語群判定は，単語の頻度の代わりに文書の関連性を考慮する．これにより，文書を代表する重要語を大きく重み付け，一般的単語を小さく重み付けることができる．

3.2 本手法の重み付け

ここで，TF-IDF法，関連単語群判定の両手法による重み付けはいずれも上限が設定されていない．このため，過度に大きな重みをもつ単語が存在すると，他の単語の重みの値が軽減され，適切な重み付けができない場合がある．

この問題に対し，本手法では重みの値を[0,1]の範囲に設定し、全単語からみた重み付けではなく，2文書間の類似性に着目する．類似した文書には似た表現が出現することが多い．文書1の単語wの重みを文書2の単語群からみた評価として表すことで，2文書間に似た表現が出現すればwの重みを動的に大きくすることができる．これにより文書間の類似性を適切に測ることが可能になる．

---

*"Person Identification System for Web Documents Based on Fuzzy Integral"

Okihisa UTSUMI†, Einoshin SUZUKI‡

†Graduate School of Engineering, The University of Tokyo

‡Division of Electrical and Computer Engineering, Faculty of Engineering, Yokohama National University

重み付けに使用するパラメータは，単語の頻度からみた珍しさと単語間の関連度である．ここで関連度とは，他の単語と共起する割合である．各単語間で関連度は相互に2つ存在する．それに各単語の珍しさと合わせて，合計4つのパラメータを使用する．

関連単語群判定では単語の頻度は考慮していない．よって，一般的単語の重みは，同じ排他性をもてば，頻度に関係なく同等になってしまう．そこで本手法では，頻度もパラメータに含むことで，頻度が多く多文書に存在する単語ほど一般性をもたせる．

このように複数の基準が存在する場合，ファジィ積分[2]を用いると最も関連する単語に適切な評価値を与えることができる．本研究では複数の属性(単語)間を考慮する場合に有効な，S積分[2]を適用した．

3.3 S積分の適用

S積分は，各単語間で4つのパラメータの最小値をとり，求まった値全てについての最大値を総合評価とする．各パラメータの値は[0,1]の範囲におさめ，値が大きいほどwの評価は高くなる．

式(1)により求まるxは，各単語の頻度に基づく一般性をあらわす．ここで，t_iは各単語の頻度を，Nは全単語数を示す．また，aは全単語の頻度の平均値である．

$$x = \log_{10}\left(\frac{t_i}{a}+1\right) \quad i = /1,2,\ldots N/ \tag{1}$$

ただし，xが0以下の場合は0とし，1以上の場合は1に設定している．文書1の単語wの珍しさhは$h(w) = 1-x$として定義する．

いま文書1と文書2の2文書を考えると，文書2中の単語w_jに対する文書1中の単語w_iの関連度$g(w_i,w_j)$は式(2)で与えられる．　ここで$D(w)$はwが出現するWeb文書集合である．

$$g(w_i,w_j) = \frac{|D(w_i)\cap D(w_j)|}{|D(w_j)|} \tag{2}$$

従ってwの重みは式(3)で定義される．

$$e(w_i) = \vee_{j=1}^{N_2}[h(w_i)\wedge h(w_j)\wedge g(w_i,w_j)\wedge g(w_j,w_i)] \tag{3}$$

ここで，$\wedge$は最小値，$\vee$は最大値をとる記号である．またN_2は文書2中で考慮する全単語数である．　w_iかw_jが一般的単語の場合は，関連度が高くても珍しさにより評価値は下がる．w_iが珍しくても関連度が低ければ最小値によりw_iの重みは小さくなる．このように，本手法では2文書間での関連性により単語の重み(評価値)を動的に変化させることができる．

	関連単語群判定	S積分法
B_1	82.6	**82.8**
B_2	84.9	**93.6**

表1: 平均特定率(%)

4 実験と結果

英語を対象としたWeb文書に，提案システムを用いて実験を行なった．検索エンジンにはAltaVista [3]を使用し，検索語にはAAAI-96の会議録に載っている著者50人の姓をランダムに選んだ．各姓の上位200件のWeb文書を用い，合計10,000文書を入力とした．人物特定率として，(4), (5)の2つの指標を用いる．B_1は非人物文書を含めた特定率，B_2は人物文書だけの特定率である．

$$B_1 = \frac{\text{正解の人物文書数} + \text{正解の非人物文書数}}{200 - \text{無効な}Web\text{文書数}} \tag{4}$$

$$B_2 = \frac{\text{正解の人物文書数}}{200 - \text{無効な}Web\text{文書数}} \tag{5}$$

表1に提案手法と関連単語群判定を比較した，平均特定率を示す．ここで平均とは，50人分の特定率の平均である．提案手法と関連単語群判定のB_1はほぼ等しいが，B_2では提案手法が高い特定率をあげている．これは2文書間の類似度に着目した重み付けにより，適切に人物文書をクラス分けできた結果と考えられる．提案手法のB_1がB_2に比べて低いのは，人物判定時の誤判断による影響が大きい．誤判断の内訳は，姓が地名の場合(60.3%)，文の最初の語をファーストネームと誤認識した場合(24.0%)，姓の存在する文全てが大文字の場合(14.7%)，その他(1.0%)となった．

5 むすび

本研究ではWeb文書を自動的に人物ごとに分ける問題において，ファジィ積分を用いたシステムを提案した．人物の判定において若干の問題はあるが，非人物文書を含まない文書群においては平均93%で人物を特定し，また既存の関連単語群判定より特定率が8.7%向上し，その有効性を示した．今後の課題として人物判定の精度の向上があげられる．例えば電子辞書を用いて，姓が文の2番目に位置するときには，その直前の単語がファーストネームを指すかどうかを判断することが考えられる．

参考文献

[1] 早川 毅, 鈴木英之進: 関連単語群判定を用いたWeb文書からの人物特定システム，SIG-FAI-9804, pp.61-68 (1999)

[2] 井上 洋: ファジィ理論の基礎, 朝倉書店 (1997)

[3] AltaVista, *http://www.altavista.com*