1E1-4

人物名に着目した二段階クラスタリングによる Web 上の同姓同名人物の分離

Two-step Clustering based on Person Names to Identify Different People with Identical Names on the Web

片岡 真一[*1] 上田 洋[*2] 村上 晴美[*1] 辰巳 昭治[*2]
Shinichi KATAOKA Hiroshi UEDA Harumi MURAKAMI Shoji TATSUMI

[*1] 大阪市立大学大学院創造都市研究科
Graduate School for Creative Cities, Osaka City University

[*2] 大阪市立大学大学院工学研究科
Graduate School of Engineering, Osaka City University

Person search is one of the most popular types of Web searches. Identifying people and classifying web pages from person search results is crucial in such work. In this paper, we propose a two-step clustering method based on person names to identify people and classify web pages. In the first step, initial clusters are generated based on names. In the second step, initial clusters are integrated into final clusters using personal information as features and hierarchical agglomerative clustering with a group average method.

1. はじめに

近年, Web 上で人名を用いた検索を行われることが多くなってきている. 人名を用いた検索の結果には, 同姓同名人物が含まれることが多い. それらをユーザ自身が一見して見分けることは非常に困難である. 本研究は, 氏名による Web 検索結果から同姓同名人物毎に Web ページを分離することを目的とする. これはクラスタリング手法の応用問題として定式化でき, どのような特徴ベクトルを作成するか, クラスタリング手法として何を用いるか, などが検討されてきている.

本研究では,「Web 上の同姓同名人物の分離には, Web ページに含まれる人物名の共起が他の特徴語と比べて特に有用である」という仮説に基づき, 人物名に着目した 2 段階のクラスタリング手法を提案する.

2. 提案手法

本研究では, 氏名を用いた Web 検索結果について同姓同名人物毎に分離するために, 2 段階のクラスタリング手法を提案する. 提案手法は, (1) 人物名に着目したクラスタリングと, (2) 特徴語ベクトルを用いたクラスタリングの 2 段階から構成される. 1 段階目で人物名に着目したクラスタリングを行って中間クラスタを作成してから, 2 段階目で中間クラスタに対して特徴語ベクトルを用いたクラスタリングを行う. 提案手法の概要を図 1 に示す.

2.1 人物名に着目したクラスタリング

氏名を用いた検索の結果得られた Web ページを, 人物名に着目して中間クラスタに分離する.

2.1.1 人物名抽出

本手法での人物名とは,「村上 晴美」のように, 姓と名で構成される文字列である.

人物名は以下のように抽出する. まず, Web ページからタグを除去する. タグを除去した Web ページに対し, Chasen を用いて形態素解析を行う. 形態素解析の結果,「名詞-固有名詞-人名-名」と判定された語と,「名詞-固有名詞-人名-姓」と判定された語が続けて出現した場合, その 2 つの語を抽出して人物名とする. 人物名の抽出は, Web ページ毎に行う.

連絡先:片岡 真一,大阪市立大学大学院創造都市研究科,
〒558-8585 大阪市住吉区杉本 3-3-138,
m07uc505@ex.media.osaka-cu.ac.jp

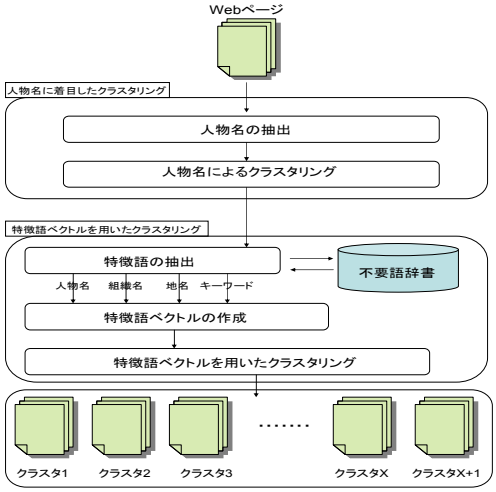

図 1 手法概要

2.1.2 人物名を用いたクラスタリング

人物名を用いたクラスタリング手法を, 図 2 に示す.

Step.1 任意のWebページを選択し, そのWebページを持つ中間クラスタを作成する.
Step.2 最初に選択したWebページに含まれる人物名と, 中間クラスタに属さないWebページに含まれる人物名を比較する. 3人以上一致する人物名を含んでいるWebページを, 作成した中間クラスタに追加する.
Step.3 中間クラスタに属さない任意のWebページを選択し, そのページを持つ新たな中間クラスタを作成し, Step.2を行う.
Step.4 全てのWebページがいずれかの中間クラスタに属するまで, Step.2とStep.3を繰り返す.

図 2 人物名を用いたクラスタリング手法

2.2 特徴語ベクトルを用いたクラスタリング

人物名に着目したクラスタリングにて作成された中間クラスタに対し, 特徴語ベクトルを用いたクラスタリングを行う.

2.2.1 特徴語ベクトルの作成

中間クラスタに含まれる Web ページから特徴語ベクトルを作成する。本手法では、以下を特徴語とする。

- 人物名
- 組織名
- 地名
- キーワード

人物名は、2.1.1 節と同じ抽出方法である。組織名、地名は、Chasen にて組織名、地名と判定された語である。キーワードは、連接する名詞をつなげたものである。なお、あらかじめ作成した不要語辞書を用い、不要語を除去する。

これらの特徴語を中間クラスタ毎に抽出、重みを計算し、特徴語ベクトルとする。特徴語の重み w は、以下のように定義する。

$$w = tf \cdot idf = tf \cdot \left(1 + \log\left(\frac{N}{df}\right)\right)$$

なお、tf は特徴語の頻度を正規化したもの、N は中間クラスタの数、df は N に対する特徴語の出現回数である。

2.2.2 クラスタリング

中間クラスタから作成した特徴語ベクトルを用いてクラスタリングを行う。クラスタリングは階層型クラスタリングを用い、距離の計算に群平均法を使用する。距離 $D(C_1,C_2)$を、

$$D(C_1,C_2) = \frac{1}{|C_1||C_2|} \sum_{x_1 \in C_1, x_2 \in C_2} d(x_1,x_2)$$

と定義する。なお、C_1, C_2 はクラスタリングにより生成されるクラスタ、x_1, x_2 は中間クラスタ、$d(x_1,x_2)$ は中間クラスタ x_1,x_2 間の類似度である。$d(x_1,x_2)$ は

$$d(x_1,x_2) = \frac{\sum_{i=1}^{T} w_{xi} w_{yi}}{\sqrt{\sum_{i=1}^{T} w_{xi}^2 \sum_{i=1}^{T} w_{yi}^2}}$$

と定義する。なお、T は得られた特徴語の総数を、w_{xi} は x_1 、w_{yi} は x_2 が持つ特徴語である。

3. 実験

本手法の有効性を確認するために予備的な実験を行った.

3.1 方法

氏名「江川 卓」,「田中 克己」,「菱沼 聖子」,「三浦 麻子」を用いて Google WEB APIs で Web 検索を行い, 上位 100 件の Web ページを取得した. 得られた Web ページを提案手法を用いて同姓同名人物毎に分離した. 閾値毎(0.0001,0.0005,0.001,0.005,0.01)に F 値による評価を行った. F 値の計算には, [Wan 05]のクラスタ評価で用いられた式を使用した.

3.2 結果と考察

平均で, 閾値 0.0001 で F 値が 0.830, 0.0005 で 0.826, 0.001 で 0.829, 0.005 で 0.864, 0.01 で 0.835 であった. 本手法では, 閾値 0.005 がもっとも良い評価であった(図 3 参照).

氏名別に最も F 値が高かった閾値は,「江川 卓」で閾値 0.005(F 値 0.809),「田中 克己」で閾値 0.01(F 値 0.939),「菱沼 聖子」で閾値 0.0001(F 値 0.885),「三浦 麻子」で閾値 0.001 F 値(0.954)であった.

F 値による評価はおおむね良好である.

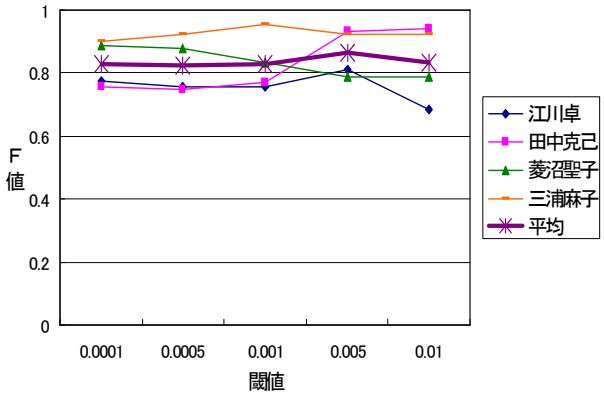

図 3 実験結果

4. 関連研究

Wan らはプロフィール情報を含む多様な特徴語ベクトルを用いたクラスタリングを行っており, 本研究における 2 段階目のクラスタリング手法に類似する[Wan 05]. また, 人物のプロフィール情報や Web のリンク構造に着目した研究などが数多く存在する([佐藤 05], [白砂 06]など). 木村らは Web ページ全体のかわりに検索エンジンのスニペットを用いて高速化を図っている[木村 06].

本研究は, Web ページ内テキストを用いており, プロフィール情報の中でも人物名に特に着目した研究である. Wan らの研究の改良研究と位置付けられる.

5. おわりに

本研究では, Web 上の同姓同名人物を分離するために, 人物名に着目した 2 段階クラスタリング手法を提案した. まず, 人物名に着目したクラスタリングを行い, その結果を特徴語ベクトルによる階層型クラスタリング手法を用いて再度クラスタリングを行った.

予備的な実験の結果は良好であり, 閾値毎の評価では, 閾値 0.005 が F 値 0.864 で最も良かった.

今後は, 氏名数を増やした実験を行うとともに, 他手法との比較実験を行う予定である.

参考文献

[Wan 05] X. Wan, J. Gao, M. Li and B. Ding, Person Resolution in Person Search Results: WebHawk, In Proceedings of CIKM'05, pp. 163-170, 2005.

[佐藤 05] 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 No. SIG 8, 2005.

[白砂 06] 白砂健一, 小山聡, 田島敬史, 田中克己, Web の構造情報とプロファイル抽出を用いたオブジェクト識別, 第 17 回データ工学ワークショップ(DEWS2006)論文集, 2C-i7, Mar. 2006.

[木村 06] 木村塁, 戸田浩之, 田中克己: 検索結果スニペットのクラスタリングによる同姓同名人物の特定, DEWS2006, 2C-i11, 2006.