Amazon 著者名検索結果の同名異人毎への自動分類

上田　洋(d06tb001@ex.media.osaka-cu.ac.jp)　　村上　晴美(harumi@media.osaka-cu.ac.jp)
大阪市立大学大学院工学研究科　　大阪市立大学大学院創造都市研究科

概要: 利用者の著者名入力に基づき Amazon 和書の著者名検索を行い、書籍のタイトルとカテゴリデータを取得し、単一パス法に基づき検索結果の同名異人を判定し、同名異人毎に検索結果を分類して表示するシステムを開発した。同名異人が著者・編者として存在する人名を検索質問として使用した実験の結果、以下のことがわかった。(1) 著作を複数持つ人物を対象とした場合、単一パス法の閾値は 0.5 が最も良かった。(2)閾値 0.5 において、クラスタ適合率 92%、分離性能 66%、著者再現率 84% であった。 (3) 著作が一つしかない人物を含めた場合、含めない場合と比べて分離性能が悪かった。

1. はじめに

図書館では著者名典拠を用いて同名異人の識別を行う仕組みがある。著者名典拠の作成や著者標目の付与は図書館員が人手で行うために、その精度は非常に高いが、人手で行うがゆえにコストも高い。またそのコスト高のために著者名典拠を導入・維持できない図書館も数多く存在する。

一方、多くの書籍検索サイトでは著者名典拠の仕組みはない。たとえば、Amazon では、検索された書誌情報の「著者名」を選択すると、著者名の文字列で検索するため、同名異人を含む別人の書籍が検索結果に含まれる。

他方、情報科学の分野においては、クラスタリング手法を用いて膨大な文書を類似した文書群に分類する研究が行われ、部分的に実用化している。我々は、著者名典拠の仕組みを持たない書籍検索システムにおいて、クラスタリング手法を用いて書籍を同名異人毎に分類できるのではないかと考えた。

本研究では、利用者の著者名入力に基づき、Amazon 和書検索を行い、検索結果の同名異人を判定し、同名異人毎に検索結果を分類して表示するシステムを開発する。同名異人が著者・編者として存在する人名を検索質問として使用し、Amazon 和書の検索結果を取得し、同名異人毎に分類する。分類した各クラスタをクラスタ適合率、著者再現率、分離性能などの尺度で評価する。

2. 手法

人間が同名異人の書籍を区別する手がかりとして、書籍のジャンルやテーマ、共著者、発行年、出版者などがあげられる。今回は、その最も基本的な方法と考えられる、書籍のジャンルやテーマを手がかりとして分類する過程を自動化することを試みる。

本研究では、Amazon.co.jp から書籍データを取得し、非階層型クラスタリングの一種である単一パス法を用いて同名異人の識別を行う。以下では、その手法について述べる。

2.1. 書籍リスト取得

Amazon.co.jp から、著者名を用いて和書検索を行い、同名異人を含む書籍リスト（ISBN）を取得する。

2.2. 書籍データ取得

取得した書籍リスト（ISBN）から、Amazon Web Service を用いて各書籍のデータを取得する。今回は、内容に基づく自動分類を目指すため、取得したデータの中から、書籍のタイトルと、書籍のカテゴリデータである BrowseNodes 内の文字列を使用する。

2.3. 分類処理

各書籍間の類似度をベクトル空間モデルを用いて算出し、その類似度を元に非階層型クラスタリングの一種である単一パス法を使用し、同名異人毎に分類する。

2.3.1. 重み付け

書籍 b_i における索引語 t_j の重み w_{ij} を

$$w_{ij} = TF_{ij} \cdot IDF_j$$

と定義した。

TF_{ij} は

$$TF_{ij} = f_{ij} \Big/ \sum_{k=1}^{M} f_{ik}$$

である。なお、f_{ij} は書籍データ b_i（2.2.節で得られた書籍のタイトルと BrowseNodes に含まれる文字列）から形態素解析システム ChaSen によって名詞と判定された語で、かつ 2 文字以上から構成される語（本手法での索引語とする）t_j の出現回数である。M は索引語 t の総数である。

また、IDF_j は

$$IDF_j = \log \frac{N}{df(t)} + 1$$

とする。$df(t)$ は索引語 t の出現する書籍の数、N は分類する書籍の総数である。

上記の手法にて作成した索引語集合を書籍ベクトルとする。

2.3.2. 書籍間の類似度

書籍間の類似度はベクトルの余弦を用いて算出する。書籍ベクトル b_j、b_k の間の類似度 $sim(b_j,b_k)$ を

$$sim(b_j, b_k) = \frac{\sum_{i=1}^{T} w_{ij} w_{ik}}{\sqrt{\sum_{i=1}^{T} w_{ij}^2} \sqrt{\sum_{i=1}^{T} w_{ik}^2}}$$

と定義した 。なお、T は索引語の総数、w_{ij}、w_{ik} は書籍ベクトル b_j、b_k に含まれる索引語 t_i の重みである。算出した類似度を元に分類を行う。

2.3.3. 同名異人の分類

書籍間の類似度をもとに分類を行う。分類は、単一パス法 (single pass method) を用いた。単一パス法は下記の手順にて行う。

1. 最初に選択した文書（本手法では書籍データ）を、最初のクラスタとする。
2. 次に選択した文書と、その時点で存在する全てのクラスタとの類似度を計算する。
3. 計算した類似度が最も高いクラスタを選び、その類似度があらかじめ設定した閾値以上であれば、そのクラスタに割り当てる。該当クラスタがなければ、新しいクラスタを作成し割り当てる。
4. 全ての文書がクラスタに割り当てられるまで繰り返す。

なお、本手法では、最初に選択する書籍を取得した書籍リストの先頭にある書籍とし、書籍リストの順番（Amazon 和書の売れている順番）に沿って実行を行う。

3. 実験

3.1. 実験 1

3.1.1. 基本方針

本手法で用いる非階層型クラスタリング手法は、データを類似するものにまとめるという特性があり、著作が複数存在する著者の分類に適している。一方で、著作が一冊しかない著者の分類は難しい。そのため、本研究では著作が 1 冊だけの著者の分類を想定しないこととし、クラスタに含まれる書籍が 1 冊の場合、それらのクラスタをひとつのクラスタとしてまとめて提示することを基本方針とした。そこで、本実験では、書籍 1 冊のクラスタ群を集計から除外することとした。

3.1.2. 方法

複数の同名異人が存在する氏名 10 語（表 1 参照）を、入力文字列として与え、作成されたクラスタについて集計を行った。

まず、Amazon.co.jp の著者完全一致検索にて得られた書籍リスト内に何人の同名異人が存在するかについて調査を行った。次に、各書籍がどの人物の書籍かを調査した。調査は、NDL-OPAC の著者名検索機能、Wikipedia、該当人物のプロフィールなどを参考に第一著者が行った。

上記の調査を元に、システムが出力した結果について評価を行った。同名異人が存在する氏名 10 語と、クラスタリングで用いる閾値をシステムの入力とした。閾値は、0.1、0.2、･･･、0.9 をそれぞれ与え、各閾値の評価を比較した。

各氏名に存在する同名異人の数は、2～17 人、取得できた書籍数 4～75 冊であった。得られた書籍について、どの人物の書籍か調査した。その結果、得られた書籍の中でどの人物か特定できないもの（著者不明書籍）が存在した。著者不明書籍が 7 の氏名に存在し、書籍の数は 1～8 冊存在した（表 1 参照）。

表1 実験に用いた氏名に関する情報

入力氏名	同名異人数	取得書籍数	著者不明書籍
西田豊明	2	16	0
江川卓	2	33	0
林寛子	2	8	1
武田英明	2	5	0
佐藤亮一	2	35	7
山田博	5	19	1
鈴木清	6	23	7
伊藤隆	7	74	5
佐々木隆	7	63	8
鈴木博	17	75	4

基本方針で述べたように、クラスタ内の書籍が 1 冊のものをまとめ、その他クラスタと扱っている。そのため、本手法では書籍数が複数あるクラスタのみを評価した。

作成されたクラスタがどの人物かを判定する基準として、クラスタ内の書籍のうち、最も書籍の数が多い人物とした。最も書籍数の多い人物が書籍数同数のため複数存在する場合は、どの人物のクラスタとも特定しないこととした。

各クラスタをどの人物のクラスタか特定した上で、下記の 6 つの尺度を用いて評価した。

(1) クラスタ適合率

クラスタ cl_i の適合率（クラスタ適合率と呼ぶ）を以下のように定義した。

$$precision_i = \frac{|cl_i(R)|}{|cl_i(N)|}$$

なお、$|cl_i(R)|$ はクラスタ cl_i に含まれる適合書籍、$|cl_i(N)|$ はクラスタ cl_i に含まれる全ての書籍である。ただし、クラスタ内が著者不明書籍のみの場合、クラスタ適合率を計算しないこととし、クラスタ内に著者不明書籍が含まれることで、特定に有効な書籍の数が 1 冊になった場合は、クラスタ適合率を 0 とする。

(2) 著者クラスタ再現率

人物 j の再現率（著者クラスタ再現率と呼ぶ）を

$$recall_j = \frac{|cl_{\max(j)}(R)|}{|C_j|}$$

と定義した。なお、$cl_{\max(j)}$ は人物 j の書籍を最も多く含むクラスタであり、$|cl_{\max(j)}(R)|$ は、そのクラスタの適合書籍の数である。また、$|C_j|$ は得られた書籍リストに含まれる人物 j の書籍数である。

(3) F 値

F 値は情報検索の評価尺度としてよく用いられるものであり、適合率と再現率の調和平均である。

$$\text{F - measure} = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

(4) 分離性能

分離性能は、作成されたクラスタ群の中で特定できた人物の数 $|identify(ps)|$ と作成されたクラスタの数 $|cl|$ の商である。

$$DP = \frac{|identify(ps)|}{|cl|}$$

(5) 著者再現率

著者再現率は、以下のように定義した。

$$AR = \frac{|identify(ps)|}{|ps|}$$

$|identify(ps)|$は作成されたクラスタ群の中で特定できた人物数、$|ps|$ は書籍リストで確認できた人物数である。ただし、実験 1 では、著作を 2 冊以上持つ人物と数とした。

(6) クラスタリング率

クラスタリング率は、分類できた書籍の割合である。

$$CR = \frac{|N| - |cl_{other}(N)|}{|N|}$$

$|N|$ は得られた書籍数、cl_{other} は書籍が 1 冊のクラスタの集合であり、$|cl_{other}(N)|$はクラスタ集合内の全ての書籍である。

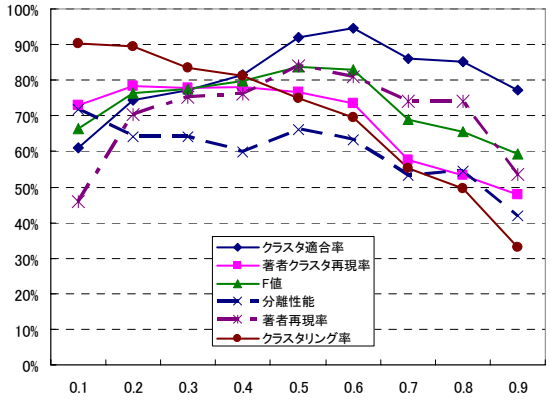

図 1 実験 1 の各閾値の結果

3.1.5. 結果と考察

集計の結果を図 1 に示す。クラスタ適合率については、0.1 から閾値を高くする毎に数値が高まり、0.6 で最高値（95%）となった。0.7 以降、クラスタ適合率が低下した原因は、書籍が複数あるクラスタが作成されない氏名が出てきたからである。著者クラスタ再現率は、閾値 0.1 から 0.6 まで 7 割を保っていたが、0.7 以降、急激に低下した。これも、クラスタ適合率と同じ原因と考える。F 値は、閾値 0.5 が最も高かった（84%）。分離性能は、0.1 の場合が最も数値がよく（72%）、次が 0.5 であった（66%）。著者再現率は、0.5 の時が最も高い（84%）。クラスタリング率については、閾値を高く設定する毎に、数値が低下する傾向にあった。

以上より、総合的に評価すると、実験 1 の条件では、閾値 0.5 で最も良い結果が得られると考える。また、閾値 0.5 での数値は、著作が複数ある著者の同名異人毎への分離に関する本手法の有効性が示唆されたものと考える。

3.2. 実験 2

実験 1 では、3.1.1.節にて述べた基本方針に則り評価を行った。その基本方針の問題点として、著作が 1 冊のみの著者が分類できないことが挙げられる。実験 2 では、著作が 1 冊のみの著者も評価の対象となるよう、実験 1 では集計を除外した書籍が 1 冊だけのクラスタについても評価を行った。

3.2.1. 方法

実験方法は、3.1.2.節で述べたものと同じであるが、実験 1 では評価の対象としなかった書籍が 1 冊のみのクラスタについても対象とした。評価尺度については、3.1.2.節の (4) 分離性能、(5) 著者再現率を用いて評価した。

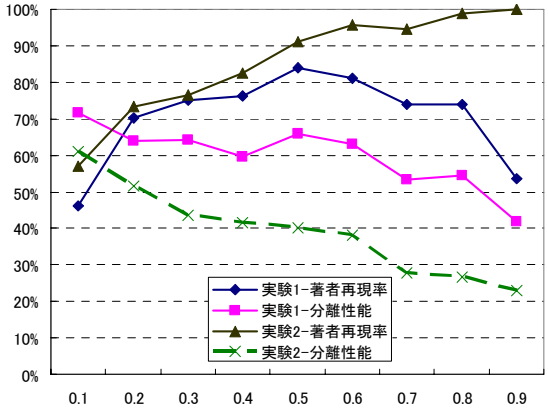

図 2 実験 2 の各閾値の結果

3.2.2. 結果と考察

図 2 に実験 2 の結果を示す。集計の結果、著者再現率は閾値を上げる毎に数値は高まり、0.5 で 9 割（91%）に達し、0.9 では 100%となった。分離性能については、閾値を上げる毎に数値は低下し、0.3 から 0.6 までは、比較的数値の下落率が抑えられているが（0.3 - 44%、0.6 - 38%）、0.7 以降 2 割台に落ち込んだ。

実験 2 では、著作が 1 冊の著者についても集計できるように書籍が 1 冊のクラスタも対象とした。著者再現率の$|ps|$にも著作が 1 冊の著者を含めた。

その結果、著者再現率については実験 1 よりも高くなっている。著者再現率の結果は、実験 1 で集計を除外したものについても対象とした結果が現れている。

対照的に、分離性能については全ての閾値で実験 1 の結果を大きく下回る結果となった。分離性能は 1 人に対してのクラスタ数の割合を表している。1 人に対してのクラスタ数が多くなると、どの人物のクラスタかを識別することが困難になる。実際にデータをみてみると、本来分類されるべきクラスタにうまく分類されなかった書籍のみからなるクラスタが出現している。

実験 2 の結果は、本手法では著作が 1 冊の著者の分類を試みた場合にあまりうまくいかないことを示唆していると考える。

表2 閾値0.5での「江川　卓」の結果

取得順位	クラスタ1
1	罪と罰〈上〉
2	罪と罰〈中〉
3	罪と罰〈下〉
4	地下室の手記
5	謎とき『カラマーゾフの兄弟』
6	悪霊（上巻）
7	悪霊（下巻）
9	謎とき『罪と罰』
11	謎とき『白痴』
15	ドストエフスキー
16	ドストエフスキー
19	ソ連潜水艦U137―人工地震エンマ作戦
23	赤いパイプライン
24	その時、その所
25	愛と詩の手紙―ボリス・パステルナーク・オリガ・フレイデンベルグ往復書簡集1910~1954
33	新潮世界文学辞典

	クラスタ2
8	江川流マウンドの心理学―野球の面白さ100倍!駆け引きバイブル
12	江川卓スカウティングレポート〈'98〉
13	プロ野球スカウティングレポート〈'97〉
14	江川卓・スカウティングレポート〈'99〉
17	たかが江川されど江川
20	たかが江川されど江川
27	スポーツうるぐす 夢野球
29	スポーツうるぐす「夢競馬」奮戦記
30	江川卓・スカウティングレポート〈2000〉
31	江川卓・スカウティングレポート〈2001〉
32	マウンドの心理学

	クラスタ3
10	世界の故事・名言・ことわざ総解説
21	世界の故事・名言・ことわざ・総解説
22	世界の故事・名言・ことわざ
26	世界の故事・名言・ことわざ 総解説

	クラスタ4
18	夢ワイン
28	夢 ワイン

実験1における「江川卓」の閾値0.5での数値：クラスタ適合率: 100%、著者クラスタ再現率: 82%、F値: 90%、分離性能: 50%、著者再現率: 100%、クラスタリング率: 100%

4. 実行例

本節では、入力氏名「江川　卓」で本手法にて得られる出力結果について解説する。この例は比較的うまく動作した例である。

「江川　卓」氏は、同名異人が2人存在する。一人は、元野球選手であり、現在は野球解説者として活躍している人物である。また、ワインにも造詣が深いことで知られている。もう一人は、ロシア文学の専門家である。前者は、「卓」を「すぐる」と読み、後者は「たく」と読むという違いはあるが、漢字が同じである同名異人として最も有名な氏名のひとつであると考える。例に閾値0.5での「江川　卓」の結果を示す（表2参照）。なお、クラスタ番号は本手法で作成された順番を示す。取得順位はAmazon.co.jpから得られた順番を表す。

出力結果は、4つのクラスタが作成される。野球解説者「江川　卓」氏のクラスタが2つ、ロシア文学専門家「江川　卓」氏のクラスタが2つ出力される。野球解説者については、クラスタ2で、主に野球に関する書籍が多く見受けられる。クラスタ4では、ワインに関する書籍が分類された。ロシア文学専門家については、クラスタ1でロシアに関する書籍がほとんどであり、クラスタ3では、世界の名言やことわざに関する解説書が分類されている。クラスタ3に関しては一見ロシア文学とは関連がないが、NDL-OPACの著者名検索での調査の結果、ロシア文学専門家の「江川　卓」氏の著作ということが判明した。

5. 関連研究

同名異人の分類に関する研究には、佐藤ら [1] や白砂ら [2] の研究などがある。佐藤らは、Web上の同姓同名人物を分類するために、階層的クラスタリングを用いている。白砂らは、Web検索結果に含まれる同姓同名人物の分類を行うために、さまざまなクラスタリングアルゴリズムやデータの特徴ベクトルでの表現方法などの既存技術のさまざまな組み合わせの比較・分析を行い、その上で、精度を向上させるために，Webの構造情報と人物に関するプロファイル抽出を用いる手法を提案している。佐藤ら、白砂らのどちらの研究もWebページを情報源として扱っているが、本研究は書籍を情報源としている点で異なっている。

6. おわりに

本研究では、利用者の著者名入力に基づき、Amazon和書検索を行い、検索結果の同名異人を判定し、同名異人毎に検索結果を分類して表示するシステムを開発した。同名異人が著者・編者として存在する人名を検索質問として使用し、Amazon和書の検索結果を取得し、同名異人毎に分類を行った。

実験の結果、以下のことがわかった。(1) 著作を複数持つ人物を対象とした場合、単一パス法の閾値は0.5が最も良かった。(2)閾値0.5において、クラスタ適合率92%、分離性能66%、著者再現率84%であった。(3) 著作が一つしかない人物を含めた場合、含めない場合と比べて分離性能が悪かった。

今後の課題としては、単一パス法以外のクラスタリングの手法との比較や、Amazonで提供されているほかの書籍データ（著者、出版社、発行年など）の使用があげられる。

参考文献

[1] 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, Vol.46, No.SIG_8, pp.26-36, 2005.6.

[2] 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, 2006.3.