



表2 第2種文の例(wは文末約物には含めない。)

ひぐらしのなく頃に と  
 鋼の錬金術師 ですかねw  
 どっちも有名ですねw  
 どっちのキャラも好きですww

表3 調査対象スレッド(\*を付したものを以外は、匿名掲示板“2ちゃんねる”のスレッド)

・キャスト！ 学生コミュニティ掲示板 アニメおたく来て！[5] (10代対象) *
・Hangame コミュニティ掲示板 自分の中のNo 1 アニメは？[6] (20代対象) *
・初めてファミコン買ってもらったときの喜びを語るスレ[7] (30代対象)
・昔は当たり前だったこと[8] (40代対象)
・50代以上アニメ好きいるかい？[9] (50代対象)
・60歳以上だけどアニメ語ろうぜ？[10] (60代対象)

## 2.2 特徴1

すべての文(第1種文および第2種文)に対する第2種文の割合は、いわゆる“お行儀の悪い文”の割合であり、その調査結果を表4に示す。10代での第2種文の割合は高いが、40代、60代にもかなりの使用例が認められる。

表4 すべての文(第1種文および第2種文)に対する第2種文の割合

年代	10	20	30	40	50	60
割合(%)	27.5	16.8	8.1	21.4	7.2	22.7

## 2.3 特徴2

表3の調査対象スレッドにおけるすべての文(第1種文および第2種文)の長さ(文字の数)の平均を表5に示す。10代から40代までは単調に増加し、それ以上の年代での違いは少ない。

表5 すべての文(第1種文および第2種文)の長さ(文字の数)の平均

年代	10	20	30	40	50	60
文字数	13.8	18.9	19.6	24.2	18.3	20.6

## 2.4 特徴3

表3の調査対象スレッドの中から100件のレスを取り出して、そこでの句読点(“ ”、”)の出現度数を調査した。その結果を表6に示す。年代と共に、句読点(“ ”、”)の出現度数は増加傾向にある。

表6 100件のレスの中での句読点(“ ”、”)の出現度数

年代	10	20	30	40	50	60
”、”	41	80	56	133	110	97
”。”	29	93	97	52	202	202
計	70	173	153	185	312	299

## 2.5 特徴4

表3の調査対象スレッドの中から100件のレスを取り出して、そこで最大行長を越える文(第1種文および第2種文)の出現度数を調査した。その結果を表7に示す。これは最大行長を意識して文の長さを短くする配慮のなさを示す指標となる。

表7 最大行長を越える文の出現度数

年代	10	20	30	40	50	60
出現度数	0	5	10	9	14	18

10代の値が小さいのは、特徴2が示すように文の長さが短いことによる。50、50代については文の長さを短くする意識は見られない。そのため、この特徴は年代を表現し易い。

## 2.6 まとめ

匿名掲示板の文体の特徴については、句読点(“ ”、”、”)の出現度数および最大行長を越える文の出現度数が、著者の年代を表現し易いことが明らかになった。文体の特徴の調査中に、文のセマンティクス(レスの内容)についても著者の年代を表現し得るものがあるが、分野への依存性が高いため、それに基づく汎用の著者年齢推定アプリケーションは構成しにくい。

## 3. 論文の文体による著者の同一性判定

### 3.1 調査対象

学術論文は、学会が用語、用字、句読点についてガイドラインを提示していることが多く、著者の特徴は顕著に顕れにくい。特に査読によってレビュー/修正された論文には、その傾向が強い。それでここでは、比較的制約の弱い講演の予稿集(画像電子学会年次大会予稿集)を取り上げて、著者30人が書いた各2件の日本語の論文(全60件)を調査対象とした。

これらの論文について、接続詞、指示代名詞、文末の表記、句読点、類義語、断言調、カタカナ語の出現を調査して、次の語の出現傾向が著者の特徴をもつことを確認した。

- (1) 順接、逆説、並列、添加、説明、選択、転換の7種類の接続詞
- (2) 指示代名詞 (こそあど言葉)
- (3) 文末の表記 (と思う、かもしれない等)
- (4) 句読点

備考：断言調、類義語、カタカナ語については、出現度数が少ないため、著者の同一性判定には不適切と判断した。

### 3.2 調査の方法と結果

同一著者が書いた2件の論文について、著者の特徴を表わす次の語を抽出し、その現れ方を調べることによって、著者の文章表記の傾向を調査した。

#### (1) 接続詞

2件の各論文から7種類の接続詞のそれぞれについて最も出現度数の高いものと2番目に出現度数の高いものを抽出する。ひとつの論文に現れるそれらの一方がもうひとつの論文にも現れるかどうかを調べ、そうであれば、傾向どおりとする。表8に、順接の接続詞の出現傾向のパターンを示す。

表8 順接の接続詞の出現傾向

接続詞	傾向の例	順接	
		調査対象著者の論文の1つ目	調査対象著者の論文の2つ目
傾向通り	類度1番目	だから	そのために
	類度2番目	したがって	だから
傾向通り	類度1番目	だから	だから
	類度2番目	したがって	その結果
傾向通り	類度1番目	だから	だから
	類度2番目	その結果	したがって
傾向通りではない	類度1番目	だから	そのために
	類度2番目	したがって	その結果
判定不可	類度1番目	だから	そのために
	類度2番目		その結果

7種類の接続詞のそれぞれについて、29人の著者に傾向通りが認められた。

#### (2) 指示代名詞

2件の各論文からすべての指示代名詞“〇れ”の出現度数と、すべての指示代名詞“〇の”の出現度数を調査する。両出現度数の大小関係が2件の論文について同じであれば、傾向どおりとする。表9に、指示代名詞の出現傾向のパターンを示す。

表9 指示代名詞の出現傾向

表10 文末の表記 傾向の例		
	調査対象著者の論文の1つ目	調査対象著者の論文の2つ目
傾向通り	この+で+は+の+と+>+に+で+れ+あ+れ+と+れ	この+で+の+は+の+と+>+に+で+れ+あ+れ+と+れ
傾向通り	この+で+の+は+の+と+<+に+で+れ+あ+れ+と+れ	この+で+の+は+の+と+<+に+で+れ+あ+れ+と+れ
傾向通りではない	こ+で+れ+あ+れ+と+れ<+に+で+れ+あ+れ+と+れ	この+で+の+は+の+と+<+に+で+れ+あ+れ+と+れ
傾向通りではない	こ+で+れ+あ+れ+と+れ>+に+で+れ+あ+れ+と+れ	この+で+の+は+の+と+>+に+で+れ+あ+れ+と+れ
判定不可	この+で+の+は+の+と+ = +に+で+れ+あ+れ+と+れ	この+で+の+は+の+と+>+に+で+れ+あ+れ+と+れ
判定不可	この+で+の+は+の+と+ = +に+で+れ+あ+れ+と+れ	この+で+の+は+の+と+ = +に+で+れ+あ+れ+と+れ

この出現傾向のパターンについては、22人の著者に傾向通りが認められた。

### (3) 文末の表記

“である”, “だ”などの文末の表記を調査する。同じ著者に関しては同じ文末表記が使われることが多いので、2件の各論文についてその傾向を調べる。表10に、同じ文末表記の出現傾向のパターンを示す。

表10 文末表記の出現傾向

文末の表記 傾向の例	調査対象著者の論文の1つ目	調査対象著者の論文の2つ目
傾向通り	と思う したい	したい
傾向通り	したい	と思う したい
傾向通り	と思う	と思う
傾向通りではない	と思う したい	かもしれない
傾向通りではない	と思う	したい
判定不可		したい

この出現傾向のパターンについては、15人の著者に傾向通りが認められた。

### (4) 句読点

句読点の組合せ“、。””, “、”, “、”, “、.”の出現を調査する。同じ著者に関しては同じ句読点の組合せが使われることが多いので、2件の各論文についてその傾向を調べる。なお、句読点の組合せについては、学会等でのルールが適用されることが多い。表11に、同じ句読点の組合せの出現傾向のパターンを示す。

表11 句読点の組合せの出現傾向

句読点 傾向の例	調査対象著者の論文の1つ目	調査対象著者の論文の2つ目
傾向通り	、 、	、 、
傾向通りではない	、 、	、 、
傾向通りではない	、 、	、 、
判定不可	、 、	、 、
判定不可	、(半角) 、(半角)	、 、

25人の著者に傾向通りが認められた。

### 3.3 著者の同一性判定の検討

上述の調査結果に基づき、著者の同一性判定のフィージビリティを検討している。判定したい本人の論文(参照論文)と調査対象の論文について、3. で用いた著者の特徴を表わす語:

- (1) 接続詞 [n = 1]
- (2) 指示代名詞 [n = 2]
- (3) 文末の表記 [n = 3]
- (4) 句読点 [n = 4]

の出現傾向を調べ、傾向通りのときそれぞれの判定結果 $J_n$  ( $n = 1, 2, 3$  or  $4$ )を1とし、傾向通りでないときそれぞれの判定結果 $J_n$ を0とする。

次に ( $J_1 + J_2 + J_3 + J_4$ ) の値を求め、それによって同一性判定を判定する。この著者同一性判定のアルゴリズムの概要を図1に示す。図2は、アルゴリズムの実装画面である。

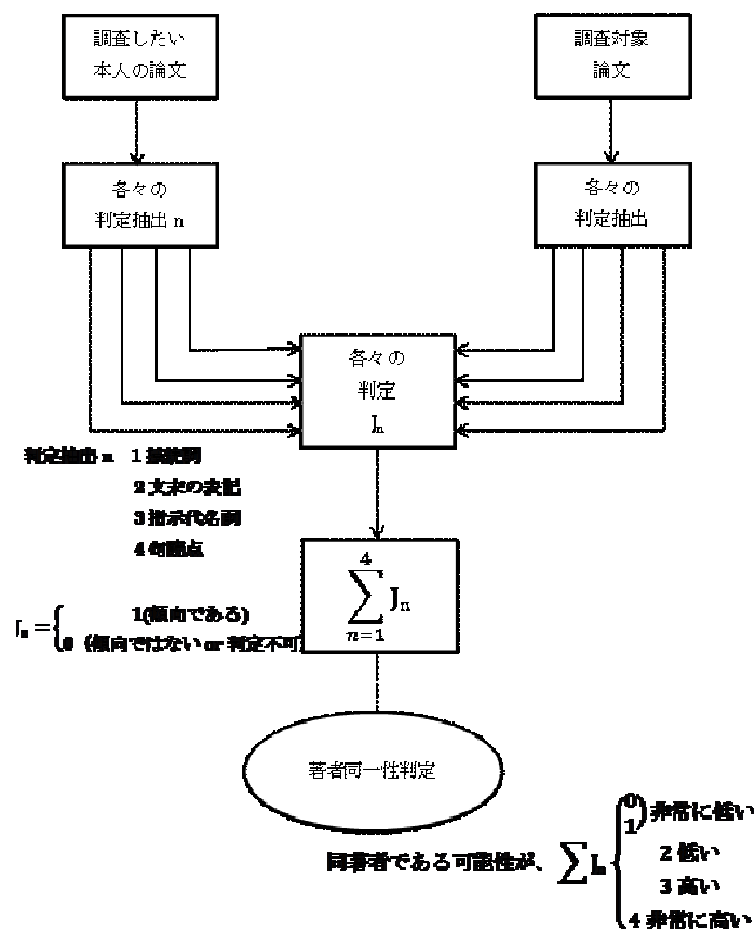


図1 著者同一性判定のアルゴリズム

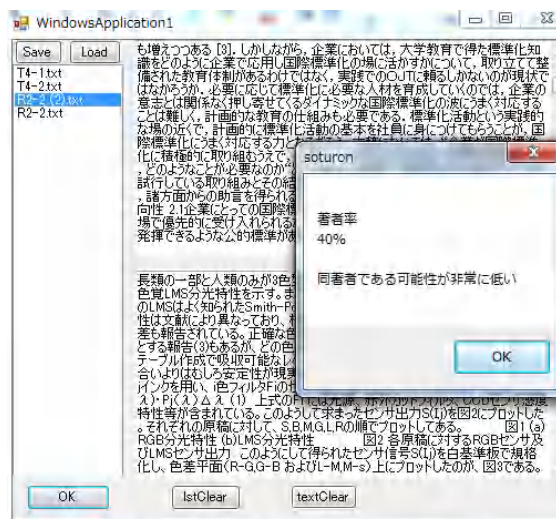


図2 著者同一性判定アルゴリズムの実装

## 4. むすび

匿名掲示板の文体の特徴については、句読点(“ ”、”)の出現度数および最大行長を越える文の出現度数が、著者の年代を表現し易いことが明らかになった。

論文の文体による著者の同一性判定については、接続詞、指示代名詞、文末の表記、句読点の出現傾向パターンが著者の特徴をもつことを確認し、その出現度数をカウント・ソートして、著者の文体の固有性を調査した。この結果に基づき、著者の同一性判定のアルゴリズムの検討を行い、その実装に着手している。

## 文献

- [1] Twitterのつぶやきから年代や性別などが分かるプロフィール自動推定技術の開発に成功” , [http://www.kddi.com/business/oyakudachi/square/labo/033/index.html?aid=bd\\_bd\\_mail110202\\_00378/](http://www.kddi.com/business/oyakudachi/square/labo/033/index.html?aid=bd_bd_mail110202_00378/)
- [2] 吉田 篤弘, 延澤 志保, 平石 智宣, 斎藤 博昭: “著者判別に有効な特徴量の推定”, 情報処理学会研究報告, 情報学基礎研究会報告, 2001(86), 83-90, 2001-09-10
- [3] 松浦 司, 金田 康正: “n-gram分布を用いた近代日本語小説文の著者推定, 情報処理学会研究報告, 自然言語処理研究会報告 99(95), 31-38, 1999-11-25
- [4] 國廣 直樹, 長谷川 智史, 穴田 一: “文章における著者の特徴解析”, 2011年電子情報通信学会総合大会, 基礎・境界講演論文集, p195
- [5] キャスフィ! 学生コミュニティ掲示板 アニメおたく来て! (何のアニメでもok♪) , <http://www.casphy.com/bbs/test/read.cgi/anime/1307960309/150>
- [6] Hangame コミュニティ掲示板 自分の中のNo 1 アニメは?, <http://forum.hangame.co.jp/thread/read.nhn?threadno=199187&page=1>
- [7] 初めてファミコン買ってもらった時の喜びを語るスレ, <http://game9.2ch.net/test/read.cgi/retro/1101044607/>
- [8] 昔は当たり前だったこと, <http://toki.2ch.net/test/read.cgi/cafe40/1311078769/>
- [9] 50代以上のアニメ好きいるかい?, <http://toki.2ch.net/test/read.cgi/cafe50/1197386510/>
- [10] 60歳以上だけけどアニメ語ろうぜ, <http://toki.2ch.net/test/read.cgi/cafe60/1305881876/150>