様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　２ 年　 ７ 月　 １ 日現在

機関番号：３４４０７
研究種目：若手研究(B)
研究期間：2015～2019
課題番号：１５Ｋ２１５１４

研究課題名（和文）授業課題レポート文書における作成者の記述特徴のモデル化と盗用発見システムへの実装

研究課題名（英文）Modeling of writing style features of authors extracted from class report documents and implementation in plagiarism detection system

研究代表者
大野　麻子（Ohno, Asako）

大阪産業大学・工学部・講師

研究者番号：９０５５０３６９

交付決定額（研究期間全体）：（直接経費）　　2,000,000円

研究成果の概要（和文）：本研究では授業課題レポート作成者の記述スタイル特徴を定量表現し，作成者認証を行うことで盗用を発見する手法を提案し，これを実装したシステムを開発した．
実際に複数の学生が作成したレポートを用いて評価実験を行った結果，記述スタイルモデルが概ね作成者認証を正しく学習できていることを確認できたが，一部誤判定が確認された．また，学習済みの記述スタイルモデルを用いて教員が目視により抽出した学生の記述特徴と同様の特徴を読み解くことに成功した．さらに，アンケート調査から本手法による学生の精神的負担軽減の可能性が示唆された．

研究成果の学術的意義や社会的意義
本研究では「授業課題レポート文書」の性質や，「授業における盗用発見」という目的に特化し，人の行う「レポート文書の作成者の推定」に倣った，独自の類似性検出手法を提案している．「同一作成者のこれまでの書き方」をもとに，対象文書の作成者認証を行うというこれまでにないアプローチは，誤判定リスクの軽減にも貢献し，盗用発見に関わる教員・学生の精神的な負担軽減にもつながることが期待される．

研究成果の概要（英文）：I proposed a new plagiarism detection method for Japanese document based on the unique viewpoint of representing an author's writing style and performing author identification and implemented the method in the plagiarism detection system for Japanese academic reports.
The method measures similarity between the same author's previous writing style and the present one. For example, if the writing style extracted from student A's report document is different from his/her previous ones, we assume that there is a possibility of plagiarism.
I confirmed high performance for plagiarism detection from the result of evaluation experiment, though it need more improvement to be used in practice. Parameters of the trained writing models provided a variety of information that can be used to differentiate authors' writing style. The method also has a potential of achieve less burdensome plagiarism detection according to the results of questionnaire investigations.

研究分野：知的学習システム

キーワード：記述スタイル特徴　作成者認証　レポート盗用発見　記述スタイルモデル　教授学習支援システム　知的学習支援システム　特徴抽出　類似性検出

１．研究開始当初の背景

大学における授業ではレポート形式の課題や試験が課されることが多い．教員は数十名から数百名もの学生のレポートを目視で確認し，評価を行う．また，近年インターネットが普及し種々の情報源へのアクセスが容易になったことにより，レポート盗用問題が深刻化している．公正な採点を行うためには盗用をもれなく発見することが必要であるが，この作業も一般に目視・手作業により行われるため，多大な時間と労力を要する．テキスト間類似性検出手法を適用し，盗用発見を自動化・支援することで教員の負担を軽減し，授業改善や学生支援に注力できるようにすることが望まれる．

レポートのような自然言語文書における類似性検出には n-gram や TF-IDF を用いることが一般的である．しかし，授業課題レポート文書には (性質 A) 論文と比べ文字数が少なく十分な特徴の抽出が難しい，(性質 B) 同一の出題テーマに対し一斉に作成されるため，レポート文書の内容が互いに類似しやすいという性質がある．よって，文書の内容に基づく既存の盗用発見手法では，偶然の一致を盗用と誤検出してしまう恐れがある．また，盗用の方法には (ケース A) 学生同士のレポート文書のコピー，(ケース B) インターネットからのコピー，(ケース C) 第三者による代筆，と様々な種類がある．特にこのうち，(ケース C) の第三者による代筆のケースでは比較対象となる文書が存在しないため，既存手法による盗用発見が難しい．

２．研究の目的

本研究ではこれまでに，プログラミング授業課題として提出されたソースコードを対象とした二種類の類似性検出手法を提案した．このうち盗用発見のために開発した「ソースコード作成者の記述特徴に基づく類似性検出手法」において，既存手法で危惧される「偶然の一致」を盗用と誤判定してしまうケースにおいても，誤判定なく正確に盗用発見を行えることを確認した．

本研究では，この「記述特徴に基づく類似性検出手法」を授業課題レポートに適用し，前述の三つのケースの盗用に対応可能な手法を提案し，これを実装した盗用発見システムを開発する．

３．研究の方法

まず，授業課題レポート文書の性質を考慮した類似性検出手法を提案し，これを実装した盗用発見システムを開発する．学生に課題を課してレポートを作成させることにより，学生の作成したレポートのセットを用意する．このセットには盗用が含まれていないことが保証されている．開発した盗用発見システムを用いて対象となるレポートのセットの作成者特徴を記述スタイルモデルに学習させ，モデルが作成者特徴を定量表現できているか確認する．学習済みのモデルの出力確率を用いて盗用発見を行い，その精度を確認する．また，盗用発見に付随する問題として挙げられる精神的負担の軽減についてもアンケート調査を通して検証する．

(1) 記述スタイル特徴に基づく授業課題レポート盗用発見手法の提案

まずは教員の目視によるレポート盗用発見の手続きとしてどのような指標や方法が用いられているか調査・分析する．このメカニズムを図 1 に示す本手法の基本的なアイディアに組み込むための検討を行う．

「基本的なアイディア」とは具体的に次の通りである：過去に提出されたレポート文書から，文字を記号とみなしたときの記号列のパターンに関する特徴を「これまでの書き方」「作成者の記述特徴」として抽出し，新たに提出されたレポート文書から抽出された特徴と照合し同一人物（正確には「同じ書き方」）により書かれたレポートであるか判定することにより盗用発見を行う．授業内外を問わず，他者が執筆した文書との類似に基づく盗用発見は行わない．

検討結果を元に，本手法のアルゴリズムを再構築し，記述スタイルモデルを完成させる．

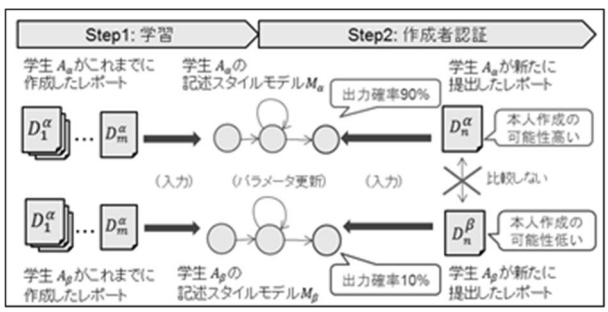

図 1　記述スタイル特徴を用いたレポート盗用発見手法の基本的なアイディア

(2) 授業課題レポート盗用発見システムの開発とモデルの評価

提案手法を実装した授業課題レポート盗用発見システムを開発する．レポート盗用発見システムは Java で開発する．システムは「前処理部」と「学習部」により構成される．「前処理部」ではレポート文書を構文解析し，本手法の定義に従いトークン化する．次に一つのレポート文書から得られたトークン列を切り出し，各モデルに入力する学習データのセットを生成する．「学習部」には隠れマルコフモデルの構造を定義してあり，入力データからモデルのパラメータを更新する．また，新たに提出されたレポートから生成した入力データから出力確率を算出する．

開発したシステムにレポートのセットを入力し，それぞれの作成者の記述特徴を学習させた記述スタイルモデルを用意する．作成者の記述特徴が正しく学習できているか，モデルを用いて作成者認証が行えるか実験を行い評価する．

(3) 盗用発見における精神的負担の軽減効果ついての検証

盗用発見は公平な成績評価に欠かせないものである一方，最終的には盗用の有無について対話による事実確認を行う必要があり，教員にとっても学生にとっても精神的に負担を感じる可能性が高い．

提案手法は他人の書いた文書との比較でなく，作成者のこれまでの書き方との比較を元にしたアプローチであるため，盗用発見を行う際の教員・学生の精神的負担を軽減できる可能性があり，先に行った予備調査では良好な結果を得ている．学生を対象としたアンケート調査を実施し，本手法に盗用発見における精神的負担を軽減する効果が期待できるか検証する．

４．研究成果

(1) 記述スタイル特徴に基づく授業課題レポート盗用発見手法の提案

先に考案した基本的なアイディアをもとに記述スタイル特徴に基づく授業課題レポート盗用発見手法を提案した．図 2 に提案手法の流れを示す．まず，学生のレポート文書をトークン化する．トークンには基準となるトークン basing-point token とその前後に出現する identification token の二つの種類がある．トークン列を先頭からチェックしてある basing-point token が現れた時，その前後に同じ basing-point token が現れるまで，または先頭・末尾のいずれかに行きつくまでトークン列上をたどり，切り出しを行う．切り出したトークン部分列が，対象となる basing-point token に関わる作成者特徴を表す記述スタイルモデルの入力データとなる．すなわち，ある basing-point token の前後にどの identification token トークンがどのようなパターンで現れるかを定量的に表したものが記述スタイル特徴である．

一つのレポート文書から生成されたトークン列から，basing-point token の数だけ異なる前処理を行い，結果として basing-point token の数だけ異なる記述スタイルモデルの入力データ群を得る．これをそれぞれのモデルに入力して記述スタイル特徴を学習させる．

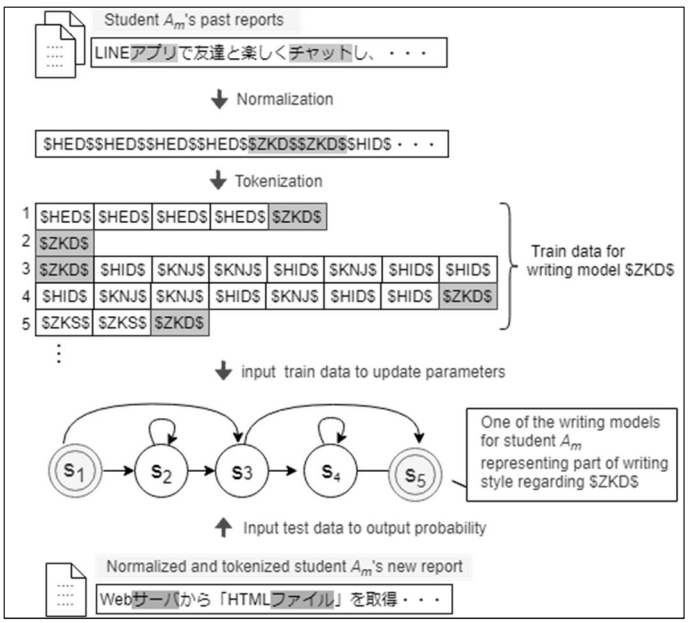

図 2　提案手法の流れ

表 1 は本手法にて定義したトークンの一覧である．レポート文書はこの変換テーブルに基づきトークン列に変換される．トークンの種類は学生の作成したレポートを対象に基礎分析を行い，作成者によらず出現頻度が高いものをリストアップした上で，作成者間の出現傾向の違いを確認しながら同じ種類の文字の半角・全角への分割やカッコ，句読点，その他記号のグループ分けを行った．「句読点（punctuations）」については basing-point token として扱う際のみ PNCというトークンで統合している．

表 1　日本語とトークンの変換テーブル

Token	Description
Letters	
KNJ	a 2-byte *kanji* (Chinese character)
HID	a 2-byte *hiragana* letter in upper case
HIS	a 2-byte *hiragana* letter in lower case
HED	a 1-byte alphabet letter in upper case
HES	a 1-byte alphabet letter in lower case
ZED	a 2-byte alphabet letter in upper case
ZES	a 2-byte alphabet letter in lower case
HKS	a 1-byte *katakana* letter in lower case
HKD	a 1-byte *katakana* letter in upper case
ZKS	a 2-byte *katakana* letter in lower case
ZKD	a 2-byte *katakana* letter in upper case
HSJ	a 1-byte number
ZSJ	a 2-byte number
Punctuations	
PNC	1-byte and 2-byte punctuation (only used as a basing point token)
HEP	a 1-byte English punctuation
HJP	a 1-byte Japanese punctuation
ZEP	a 2-byte English punctuation
ZJP	a 2-byte Japanese punctuation

Symbols	
HBC	a 1-byte bracket
ZBC	a 2-byte bracket
RTN	a linefeed
HSP	a 1-byte space
ZSP	a 2-byte space
TAB	a tab
HKG	a 1-byte symbol
ZKG	a 2-byte symbol
Others	
OTH	a special token inserted as a delimiter

本手法では基礎分析の結果をもとに，表 2 に示す 3 種類の basing point token を定義した．すなわち，3 種類の記述スタイルモデルを用いて一人の学生の記述スタイル特徴を表す．

表 2　Basing point token の例

Token	Description	Share
KNJ	a 2-byte *kanji* (Chinese character)	50 – 60 %
HID	a 2-byte *hiragana* letter in upper case	25 – 35 %
PNC	a 2-byte punctuation (ZEP and ZJP)	2 – 5%

(2) 授業課題レポート盗用発見システムの開発とモデルの評価

図 3 は学習済みの記述スタイルモデルの一例である．漢字の前後における記述スタイル特徴を表す．これに加え，ひらがな前後，句読点前後の記述スタイル特徴を表すモデルを全て用いて一人の作成者の記述スタイル特徴を表す．

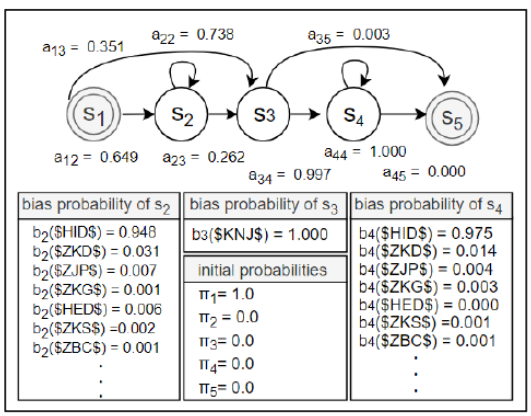

図 3 学習後の記述スタイルモデルの例

表 3　類似度（記述特徴の距離）

	M_1^{all}	M_2^{all}	M_3^{all}	M_4^{ave}	M_5^{ave}	M_6^{ave}
d_1^{ave}	0.01549	0.02748	0.03532	0.01823	0.01884	0.02054
d_2^{ave}	0.03007	0.02022	0.03568	0.02753	0.02907	0.01768
d_3^{ave}	0.04263	0.03973	0.02049	0.03074	0.02842	0.02650
d_4^{ave}	0.02098	0.03327	0.02453	0.01609	0.01563	0.01568
d_5^{ave}	0.01716	0.03723	0.02587	0.01602	0.01596	0.01798
d_6^{ave}	0.02750	0.03534	0.03014	0.02105	0.02473	0.02080

表 3 は三つのモデルが表す記述スタイル特徴と新たに提出されたレポート文書から抽出された記述スタイル特徴の非類似度（距離）を示す．作成者 5 および 6 のモデルが本来の作成者の執筆したレポートを正しく識別できなかったが，それ以外については全て正しく作成者認証を行うことができた．ただし，作成者認証を正しく行えなかったケースでも，三つのモデル全てではなく一部を用いた場合には正しく作成者認証できた．検証の結果から，現時点で採用している三つのモデルに追加，入れ替えを行うことや，レポート単位（全ての入力系列に対する出力確率の平均値を用いるの）ではなく部分系列単位（個々の入力に対する出力確率）で作成者認証を行うことで精度の向上が見込まれる．

また，学習済みのモデルのパラメータを確認し作成者の記述スタイル特徴を表現できているか検証した．表 4 には図 3 で示したモデル上の状態s_2における identification token の観測確率の抜粋，表 5 には図 3 のモデルの各状態間の遷移確率を示している．表 4 より，例えば，作成者 1（A_1）の記述スタイル特徴では漢字の前にひらがなが，作成者 6（A_6）の場合はカタカナが出現することが多い．また，作成者 2（A_2）は決して漢字の直前に「、」「。」を入れないことが分かる．表 5 からは，作成者 1 以外は漢字の後に必ず漢字以外の文字が現れ，作成者 4 と 6 以外は必ずそれが 2 文字以上続く．作成者 1〜3 に比べ 4〜6 は漢字の前に別の文字が現れることが少なく，かつ現れた文字が連続することも少ないことが読み取れる．

前述の例のような情報を三つのモデルから個別に読み取り，作成者固有のプロファイルとしてまとめることで，記述スタイル特徴を表現することに成功した．この内容はレポート文書を目視で確認した結果とも一致した．しかしこれらのモデルから情報を読みとき，プロファイル化するには専門知識と労力が必要となるため，記述スタイル特徴を説明する可読性の高いモデルの構築が今後の課題の一つとして，挙げられる．

表 4　s_2における観測確率（抜粋）

A_m	b_2(\$HID\$)	b_2(\$ZKD\$)	b_2(\$ZJP\$)
A_1	0.948	0.031	0.007
A_2	0.862	0.087	0.000
A_3	0.867	0.093	0.023
A_4	0.764	0.167	0.026
A_5	0.803	0.096	0.037
A_6	0.503	0.328	0.027

表 5　遷移確率

A_m	$a_{1,2}$	$a_{2,2}$	$a_{3,4}$	$a_{4,4}$
A_1	0.649	0.738	0.997	1.000
A_2	0.545	0.731	1.000	1.000
A_3	0.582	0.706	1.000	1.000
A_4	0.352	0.525	1.000	0.998
A_5	0.367	0.504	1.000	1.000
A_6	0.395	0.627	1.000	0.999

(3) 盗用発見における精神的負担の軽減効果ついての検証

本手法の副次的効果として，盗用発見において学生が受ける精神的負担が軽減される可能性について検証した．大学・短大の学生 269 名を対象に行ったアンケート調査の結果を図 4 に示す．上の”Method A”が本手法，”Method B”が既存手法である．本手法では「今までの書き方と今回の書き方が違う」場合に，既存手法では「他の学生の書き方と似ている」場合に教員が学生に対し盗用か否かの確認を行う．

これについて「(1)精神的負担を感じる」と回答した学生は本手法では 24.2%であるのに対し，既存手法では 46.1%と 2 倍近い値になっており，有意差が確認された（t=-6.57, df = 268, p < 0.001）．以上の結果から，本手法のアプローチが従来のアプローチに対しより精神的負担が低いものである可能性が示唆された．

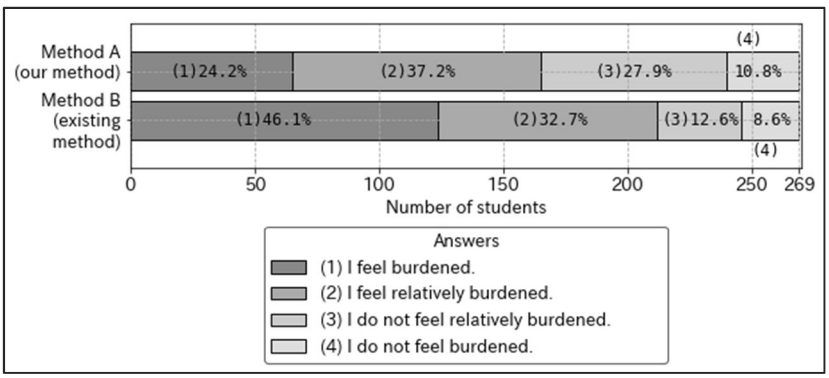

図 4　学生の精神的負担に関するアンケート調査の結果（A: 本手法，B: 既存手法）

５．主な発表論文等

〔雑誌論文〕　計4件（うち査読付論文　4件／うち国際共著　1件／うちオープンアクセス　0件）

１．著者名 Asako Ohno, Yoshihiro Ohata, Takahiro Yamasaki, and Kin-Ichiroh Tokiwa	４．巻 Volume 5, Issue 3-4
２．論文標題 Could a Customer's Migratory Behaviour in Inner Areas Explain his/her Purchase: an Exploratory Analysis	５．発行年 2016年
３．雑誌名 International Journal of Computational Intelligence Studies	６．最初と最後の頁 303,316
掲載論文のDOI（デジタルオブジェクト識別子） https://doi.org/10.1504/IJCISTUDIES.2016.083576	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Shunsuke Doi, Yoshiro Imai, Koji Kagawa, Asako Ohno, Primoz Podrzaj, Tetsuo Hattori	４．巻 第139巻, 第11号
２．論文標題 Proposal and Development of Web-based Programming Educational System with Error Analysis and Visualization	５．発行年 2019年
３．雑誌名 電気学会論文誌C	６．最初と最後の頁 1241,1247
掲載論文のDOI（デジタルオブジェクト識別子） https://doi.org/10.1541/ieejeiss.139.1241	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

１．著者名 Asako Ohno, Takahiro Yamasaki, Kin-ichiroh Tokiwa	４．巻 第140巻, 第2号
２．論文標題 Similarity Measurement Based on Author's Writing Styles for Academic Report Plagiarism Detection	５．発行年 2020年
３．雑誌名 電気学会論文誌C	６．最初と最後の頁 235,241
掲載論文のDOI（デジタルオブジェクト識別子） https://doi.org/10.1541/ieejeiss.140.235	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Asako Ohno	４．巻 vol.6, no.5
２．論文標題 Application of Content-and-Style Based Source Code Similarity Measuring Methods towards Programming Education Support System	５．発行年 2015年
３．雑誌名 ICIC Express Letters, Part B: Applications	６．最初と最後の頁 1405,1410
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計4件（うち招待講演　0件／うち国際学会　2件）

１．発表者名
Asako Ohno, Takahiro Yamasaki, Kin-ichiroh Tokiwa, and Kazuhide Togai

２．発表標題
Modeling of Authors' Writing Styles to Detect Plagiarism in Japanese Academic Reports

３．学会等名
Proceedings of the Fourth International Conference on Electronics and Software Science（国際学会）

４．発表年
2018年

１．発表者名
Asako Ohno, Takahiro Yamasaki, and Kin-ichiroh Tokiwa

２．発表標題
Development of Programming Education Material for Elementary Students to learn Computational Thinking

３．学会等名
Proceedings of the 13th International Technology, Education and Development Conference（国際学会）

４．発表年
2019年

１．発表者名
大畑善裕，大野麻子，山崎高弘，常盤欣一朗

２．発表標題
小売店の内側エリアにおける顧客の行動パターンと購買額に関する分析

３．学会等名
第14回情報科学技術フォーラム(FIT2015)

４．発表年
2015年

１．発表者名
東　明翔，山崎高弘，大野麻子，常盤欣一朗

２．発表標題
テキストマイニングを用いた金融時系列変化点の要因分析

３．学会等名
第14回情報科学技術フォーラム(FIT2015)

４．発表年
2015年

〔図書〕　計1件

１．著者名 栂井一英，大野麻子	４．発行年 2019年
２．出版社 技術情報協会	５．総ページ数 512
３．書名 「第8章第1節 運転習熟過程のモデル構築とその自律運転への適用」，車載HMIの開発動向と自動運転，ADASへの応用	

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考