様式 C-19

科学研究費補助金研究成果報告書

平成　23 年　6 月 20 日現在

機関番号：44421
研究種目：研究活動スタート支援
研究期間：2009 ～ 2010
課題番号：21800085
研究課題名（和文）プログラミング授業課題ソースコードにおける盗用発見システムの開発と有効性の検証
研究課題名（英文）Development and Evaluation of a Source Code Plagiarism Detection System for Programming Classes
研究代表者
大野　麻子（OHNO ASAKO）
四條畷学園短期大学・ライフデザイン総合学科・講師
研究者番号：90550369

研究成果の概要（和文）：
授業課題ソースコードにおける盗用発見を、作成者の記述特徴の類似性を用いて行うシステムを開発し、手法の改良を重ね、より精度の高い作成者認識が行えることを確認した。また、本手法を従来手法の出力の評価に用いることで強力かつ誤検出の少ない盗用発見を行えることを明らかにした。さらに、学生に対するアンケート調査やプログラミング授業担当教員との議論を通し、本手法が学生や教員にとって、比較的精神的不安を感じにくい可能性が高いことを確認した。

研究成果の概要（英文）：
I developed a plagiarism detection system specialized for programming classes to reduce instructors' physical and psychological burdens. My method utilizes features based on authors' coding styles. I had made several improvements and evaluated the system and obtained positive results. I proposed a new methodology that used my method to evaluate outputs of the existing methods based on algorithmic similarity to prevent misjudgments. The result of a questionnaire investigation and the findings from discussions with teachers suggested a high possibility that our method placed less psychological burdens to both instructors and students.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
21 年度	1,080,000	324,000	1,404,000
22 年度	660,000	198,000	858,000
年度			
年度			
年度			
総　計	1,740,000	522,000	2,262,000

研究分野：
科研費の分科・細目：
キーワード：教育工学、ソースコード盗用発見、モデル化、教育心理学、ソフトウェア開発効率化・安定化

１．研究開始当初の背景

「授業課題ソースコード盗用問題」は近年、世界的に注目が集まっている課題の一つであり、多くの盗用発見手法が提案されている。図１に示すように、既存手法は「①他の学生のソースコードと比較する手法（Prechelt ら、

02)」、「②インターネット上で盗用元ソースコードを検索する手法（Niezgoda ら、06)」「③編集ログを残す専用のエディタなどを使用させる手法（Vamplew ら、05)」に大別される。①は外部からの盗用に、②はクラス内での盗用にそれぞれ対応できず、ソースコードの類似を盗用と判定するため誤判定の可能性が危惧されている(Engels ら、07)。③は両方のタイプの盗用に対応できるが専用のシステムを教員・学生用端末にインストールする必要があるため導入が難しい。本手法はこれらのいずれにも属さない「記述特徴による作成者認証」という新しいアプローチを取っている。他者のソースコードとの比較を行わないので先に述べた誤判定の心配が無く、両方のタイプの盗用に対応可能である。また、システムのインストールは教員用端末のみであるため導入もしやすいという利点がある。

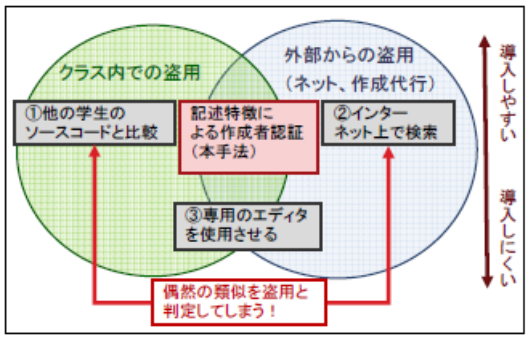

図 1 当該研究分野における本手法の位置づけ

２．研究の目的

本研究の目的は、プログラミング授業課題ソースコードから作成者の記述スタイル特徴を抽出し、これを用いて盗用発見をはじめとした授業支援を行い、教員の肉体的・精神的負担を軽減することで、間接的にプログラミング授業の質を向上させることである。

３．研究の方法

本研究の目的を達成するため、次に挙げる各項目を行った。

1. 本研究の用いる記述特徴に基づく類似性検出手法の改良
2. 本手法のシステムへの実装と評価
3. 既存システムとの比較
4. 定性的調査による教員・学生への精神的負担軽減可能性の検証
5. 授業での実際の適用に向けたシステムの改良と評価

また、システムの正確な評価を行うため、授業課題として提出されたソースコードに見立てた 240 個のサンプルソースコード集合を用意し、評価実験等に用いた。これらのソースコードは共通の出題に基づき作成されたものであり、実際の授業課題ソースコードに限りなく近い。しかし、確実に盗用が行われていないことが保証されているため、システムの出力により盗用が検出された場合はすなわち誤検出であると判断される。

４．研究成果

本研究ではソースコードの内容に基づく特徴ではなく、作成者の記述特徴に着目し、これを隠れマルコフモデルを元にした記述スタイルモデルで表現し、作成者認証を行う手法を提案した（〔雑誌論文〕②、〔学会発表〕⑥)。

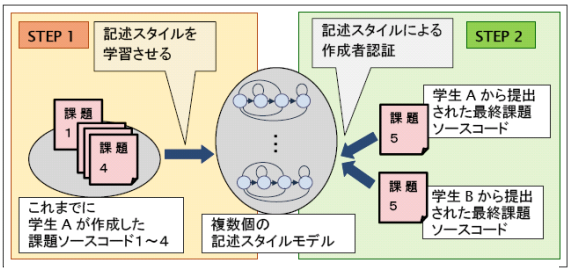

図 2 本手法の概要

図 2 は本手法の概要である。本手法では学生がプログラミング授業においてこれまでに作成したソースコードを複数個の記述スタイルモデルに学習させ、新たに提出された最終課題ソースコードから抽出した記述スタイルとモデルの表現する記述スタイルを比較することにより作成者認証を行う。

本手法では、ソースコードを記号列化し、特定の記号の前後において他の記号の出現傾向を隠れマルコフモデルをベースとした記述スタイルモデルに学習させ、記述特徴を定量化する。

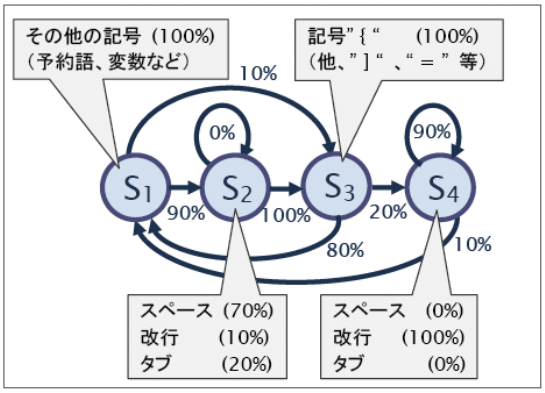

図 3 記述スタイルモデル

図 3 は記述スタイルモデルの例である。あらかじめ規定された”｛“などの記号の前後に「スペース」、「タブ」、「改行」のいずれかが出現する確率をもとに記述スタイル特徴を表現している。例えば、「学生 A のソースコードにおいて、記号”｛“の直前がスペース 1 つ、直後が改行 2 つ以上の系列が観測さ

れる確率は 56.7%」、というような特徴が個々のモデルにより表される。

このような独自のアプローチに基づく本手法の類似度算出やモデルの構造について、より高い精度で作成者認証を実現するため(1)～(3)に示すような改良を行い、(4)に示すような精神的負担の軽減に関する調査を行った。

(1) 複数の記述スタイルモデルから得られた出力確率を正規化してベクトルの要素としたものを記述スタイル特徴量とし、学習に用いたデータを入力した時に得られた記述スタイル特徴量（正解出力）との距離を求め、これを記述スタイルに基づく類似性検出結果として用いた（〔学会発表〕⑤）。

	source code by student A	source code by student B	source code by student C	source code by student D	source code by student E
s.c. by student A	100.0%	**74.5%**	51.5%	52.5%	**81.0%**

図 4 従来手法(SIM)の内容に基づく類似性検出結果

図 4 は従来の内容に基づく類似性検出手法を用いて学生 A のソースコードを学生 A～E のソースコードと比較した結果である。学生 B および C について、それぞれ 74.5%、81.0%と高い類似度が検出されている。実験に用いたソースコードには盗用がないことが保証されているため、この類似度はアルゴリズムの類似や表記方法の偶然の類似によるものと考えられる。

	models of student A	models of student B	models of student C	models of student D	models of student E
s.c. by student A	**0.1149**	0.3770	0.2693	0.3233	0.1998

図 5 本手法の記述スタイルに基づく類似性検出

図 5 は改良後の本手法により作成者特徴に基づく類似性検出を行った結果の一部であり、表中の数値は学生 A のソースコードに含まれる記述スタイル特徴量と学生 A～E の記述スタイルモデルの正解出力との距離である。で学生 A のソースコードとの類似が指摘された学生 B・C のソースコードはそれぞれ学生 A のソースコードに比べ正解出力との距離が大きいことが確認できる。また、学生 E の提出したソースコードから得られた記述特徴が学生 A の記述特徴に比較的近いことも確認できる。このように、本手法では授業課題ソースコードが本人によって作成された可能性と記述特徴の類似性を定量表現することが可能なため、従来の手法では得られなかった情報を教員に提供し、よりフェアな盗用発見を実現することができる（[雑誌論文]①、学会発表）。

(2) 記述スタイルをより詳細に表現するため、前後方記述スタイルモデルを開発した。この改良をシステムに実装し、先に作成した盗用の含まれない授業課題ソースコードを用いて、システムの評価（〔学会発表〕④）および既存手法との比較を行った（〔学会発表〕①）。

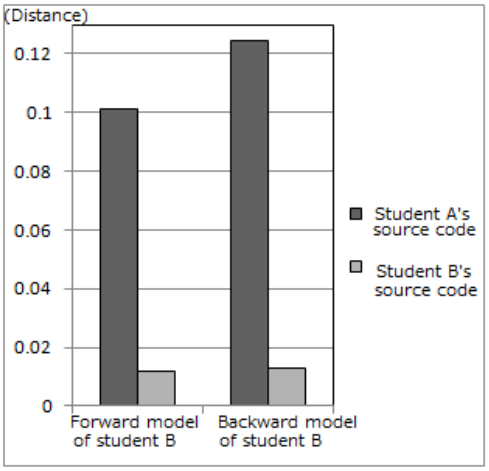

図 6 前後方記述スタイルモデルの作成者判別結果

図 6 は学生 B の前後方記述スタイルモデルに学生 A および学生 B のソースコードを入力したときの前後方記述スタイルの出力と学生 B の正解出力との距離である。濃い灰色が学生 A、薄い灰色が学生 B であり、モデルが学生 B のソースコードを A よりもはるかに近いと認識していることが読み取れる。また、前方（左）と後方（右）を比較すると、A と B の記述スタイルは後方により大きな差異があることがわかる。

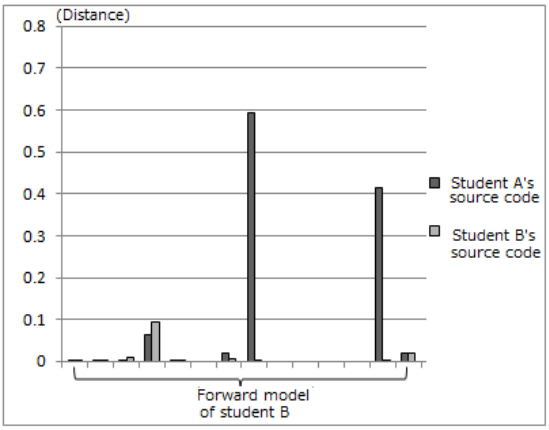

図 7 前方記述スタイルモデルの出力（詳細表示）

図 7 は図 6 の左側の前方記述スタイルモデルの出力を詳細表示したものである。これにより、教員はどの記号の後方に記述スタイルの違いが見られたか確認することができる。

(3) (2)で示したようなグラフ表示は数値に比べ記述スタイルの違いを一瞥できるメリットがあるが、GUI 上に表示するには大きな

スペースを要する。また、内容に基づく手法により高い類似性が検出されたペアが増えると、該当する学生の記述スタイルの類似を確認する作業も煩雑になっていく。そのため、より多くの学生の記述スタイルを一度に比較できるような出力表現についても検討を行った。具体的には、自己組織化マップを用いて記述スタイルの類似を元にクラスタリングを行った（〔学会発表〕③)。

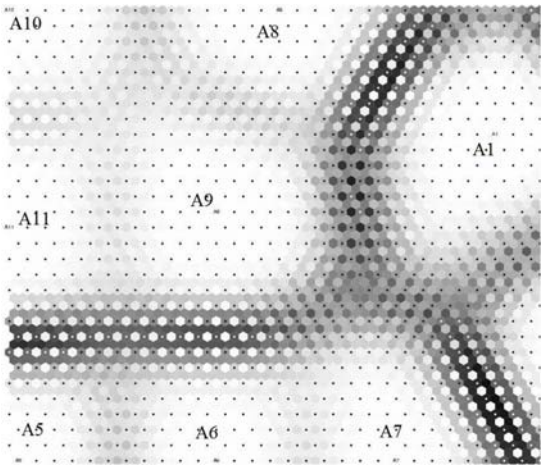

図 8 19 個のソースコードから得られた記述スタイルの分類結果

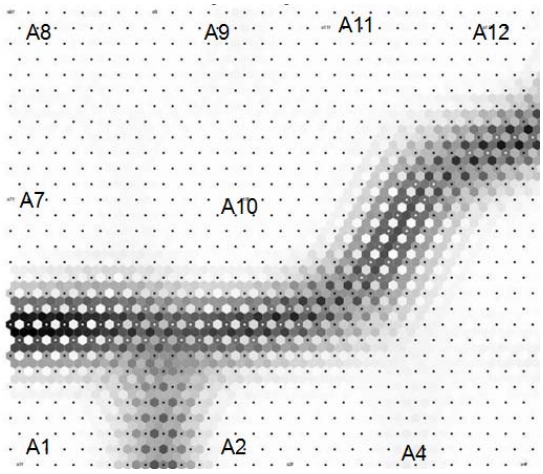

図 9 最終課題ソースコードから得られた記述スタイルの分類結果

図 8 は A1～A20 の 20 名の学生がそれぞれ作成した 19 個のソースコードを用いて記述スタイルを学習させた前後方記述スタイルモデルのパラメータを元に特徴ベクトルを生成し、自己組織化マップにより 2 次元平面上にマッピングした結果である。目視によるソースコードの確認を行った結果、近傍に配置された記述スタイル間に複数の類似した傾向が見られた。

また、図 9 に示す学生 A1～A20 の作成した最終課題ソースコードの分類結果をみると、図 8 において近傍に配された A8, A9, A10, A11 が図 9 においても近傍に配されていることがわかる。しかし、全ての分類結果について、目視による記述スタイル類似確認結果と一致していないことも事実である。今後特徴ベクトルの要素の重みづけ方法等の改良を行い、複数の記述スタイルの類似関係について高精度の分類が実現可能となれば、このような視覚的な表現を GUI に搭載する試みを行う予定である。

(4) 本手法の最大の特徴は、ソースコードの内容ではなく、作成者の記述特徴の類似度比較を行っている点である。これにより、従来の手法で危惧された偶然の一致による誤検出のリスクを軽減するだけではなく、教員が盗用発見システムの出力結果を元に盗用と判断し、学生に確認を行う際に生じる精神的負担の軽減についても効果が期待できる。

そこで、本学の 89 名の学生に盗用発見方法と精神的負担の度合いについてアンケートを実施し、既存手法と本手法による学生の精神的負担の程度に差があるか調査した（〔学会発表〕②)。

設問はおよび選択肢は次の通りである。

[設問]

授業課題ソースコードの採点後、教員から呼び出されたと想定した場合、次の A・B の質問についてどのように感じるかを 4 つの選択肢から 1 つだけ選択させた。

- 質問 A： 「あなたが提出した答案と A さん（別の学生）の答案の内容がとても良く似ていますがなぜですか？」
- 質問 B：「あなたが今回提出した答案の書き方が、今までに提出した答案の書き方とだいぶ違いますがなぜですか？」

[選択肢]

1) 精神的負担を感じると思う
2) どちらかといえば精神的負担を感じると思う
3) どちらかといえば精神的負担は感じないと思う
4) 精神的負担は感じないと思う

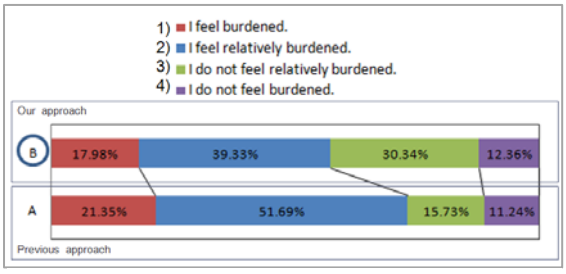

図 10 盗用発見手法と精神的負担に関する調査結果

図 10 はアンケートの集計結果である。図中上部の B の棒グラフは質問 B（本手法のアプローチ)、A は質問 A（既存手法のアプロ

ーチ）に対する学生の精神的負担の度合いを表している。「精神的負担を感じる（感じやすい）」とする 1)2)を選択した学生の割合はAに多く、「精神的負担を感じない（感じにくい）」とする3)4)を選択した学生はBに多い。このことから、本手法のアプローチが既存手法に比べ学生に精神的負担を与えにくいことが確認された。

また、本調査の結果を国際会議において発表した際、プログラミング授業を担当する教員を含む複数の教員間でディスカッションを行った結果、教員にとっても精神的負担が低いのではないかとの意見が大半であった。

（まとめと今後の課題）

プログラミング授業に関わらず、盗用は教育機関においてフェアな採点を妨げる大きな問題であり、決して放置することは許されない。しかし、その反面盗用の定義や判定基準はいまだ統一されておらず、盗用発見を自動化するツールにも一長一短があることから、現状では各教員がそれぞれに盗用の定義を行い、盗用発見を行っている。
このような中で、本研究の当該分野における貢献は、次の3点である。

① 授業課題ソースコードの持つ特徴、すなわち、(1)同じ出題意図に基づき作成されるため、アルゴリズムが類似しやすく、(2)ソースコード長が短いため、特徴が抽出しにくいという点に着目し、内容に基づく特徴ではなく作成者の記述特徴に基づく特徴量を盗用発見に用いた点
② 隠れマルコフモデルを元に開発・改良を重ねた前後方記述スタイルモデルにより、複雑性を持つ記述スタイル特徴を定量表現することに成功した点
③ 盗用発見に伴う精神的負担にも着目し、本手法の斬新なアプローチがこれを軽減させる可能性について示した点

本研究では今後も当該分野における本研究の位置づけを意識しながら本手法の改良を重ね、盗用発見システムの実用化を目指す。また、盗用発見により教員や学生が受ける精神的負担の軽減についても、大規模なアンケート調査やヒヤリングの結果を元に詳細な検証を行いたい。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計2件）

① Asako Ohno and Hajime Murao, ″ A Two-Step In-Class Source Code Plagiarism Detection Method Utilizing Improved CM Algorithm and SIM ″, International Journal of Innovative Computing, Information and Control, vol.7, no.7, (採録決定済), 査読有.

② Asako Ohno and Hajime Murao, ″A New Similarity Measure for In-class Source Code Plagiarism Detection″, International Journal of Innovative Computing, Information and Control, vol.5, no.11(B), pp.4237-4247, 2009, 査読有.

〔学会発表〕（計6件）

① Asako Ohno and Hajime Murao, ″An Author Identification of In-Class Source Codes by using the Forward-Backward Coding Models″, Proceedings of the 5th International Conference on Innovative Computing, Information and Control (ICICIC2010), Xi'an, China, December 20-22, pp. 453-458, 2010, 査読有.

② Asako Ohno and Hajime Murao, ″Work in Progress - A Novel Methodology to Reduce Instructors' and Students' Psychological Burdens in Source Code Plagiarism Detection″, Proceedings of the 40th Annual Frontiers in Education Conference (FIE2010), Virginia, U.S., October 27-30, pp.S3D-1-S3D-2, 2010, 査読有.

③ 大野麻子, 村尾　元「前後方記述スタイルに基づいた授業課題ソースコードのクラスタリング」, 『2010年度人工知能学会全国大会(第24回)論文集』(CD-ROM), 2010, 査読無.

④ 大野麻子, 村尾　元「前後方記述スタイルモデルによる授業課題ソースコード作成者特徴の抽出」, 『第37回知能システムシンポジウム資料』, pp.99-104, 2010, 査読無.

⑤ Asako Ohno and Hajime Murao, ″Modeling and Quantification of Superficial Features Extracted from Source Codes: In Consideration of Fluctuation of Description among Learning Data″, Proceedings of the 4th International Conference on Innovative Computing, Information and Control (ICICIC2009), Kaohsiung, Taiwan, December 7-9, pp.1427-1430, 2009, 査読有.

⑥ 大野麻子, 村尾　元「確率モデルによるソースコード記述スタイルの識別」, 『第 53 回システム制御情報学会研究発表講演会』, pp. 379-380, 2009, 査読無.

６．研究組織

(1)研究代表者

大野　麻子（OHNO ASAKO）

四條畷学園短期大学・ライフデザイン総合学科・講師

研究者番号：90550369

(2)研究分担者

（　　）

研究者番号：

(3)連携研究者

（　　）

研究者番号：