様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２４年　５月１５日現在

機関番号：14501
研究種目：基盤研究（C）
研究期間：2009～2011
課題番号：21500140
研究課題名（和文）：プログラム作成者モデルに基づくプログラミング授業向け盗用発見システムの構築

研究課題名（英文）：Development of a Plagiarisms Detecting System for Programming Courses based on the Programming Style Model

研究代表者
村尾　元（MURAO HAJIME）
神戸大学・大学院国際文化学研究科・准教授
研究者番号：70273761

研究成果の概要（和文）：
本研究では，プログラミング授業におけるソースコードの盗用を発見するプロトタイプ・システムを構築した．システムは，あらかじめ作成者が明らかな複数のソースコードを用いて，その「表面的な特徴」からプログラム作成者の記述モデルを作成する．そして，新しいソースコードが得られる度に，そのソースコードの作成者とされる人物の記述モデルと照らし合わせ，ソースコードが実際に記述モデルの対象となっている作成者が書いたものであるかどうかを判定する．研究室の学生を対象とした演習で作成されたプログラムを用いて，プログラムの作成者を正しく認識できるかどうかテストを行ったところ，80%以上の正解率が得られた．

研究成果の概要（英文）：
In this study, we have developed a prototype system to detect source-code plagiarisms in programming courses. The system firstly constructs “programming style models” for each of students from “expression feature” of source-codes, such as a number of spaces after some specific characters like comma, the location of new-lines around braces, etc. Then, whenever a student submits a new source-code, the system uses the student’s style model to validate if the source-code has been actually written by the student. We have tested the system in our laboratory and observed quite fair result over 80% of correct classification.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	1,200,000	360,000	1,560,000
2010 年度	1,200,000	360,000	1,560,000
2011 年度	800,000	240,000	1,040,000
年度			
年度			
総　計	3,200,000	960,000	4,160,000

研究分野：総合領域
科研費の分科・細目：情報学・知能情報学
キーワード：学習と知識獲得，隠れマルコフモデル，リファクタリング

１．研究開始当初の背景

日本の産業界におけるコンピュータ技術者不足は慢性的となっており，いわゆる文系学部出身者が技術者として IT 系企業へ就職する例も多くなっている．これにともない，申請者が属するような，いわゆる文系学部においてもプログラミングの授業が行われるようになってきている．

プログラミングの授業は実習形式で行われる場合が多いが，この手の授業の例外に漏れず，学生間におけるコピーすなわち盗用が常に問題となっている．とは言え，提出された大量のソースコードを対象とした盗用発見の試みは，採点に携わる教員への負担が大い．したがって，適正なプログラミング教育を実施するため，また，それに関わる教員の負担軽減のためにも，コンピュータによる支援が必要不可欠である．

一方，社会的，経済的な影響も大きいことから，コンピュータによるソースコードの盗用発見については，これまでも様々な研究がなされている．それらは大きく 1）文字列マッチングに基づく方法，2）プログラムの構造に基づく方法，3）意味に基づく方法の 3 つに大別できる．しかし，そのいずれもが，アルゴリズムの同一性について調べるものである．

知的財産保護の観点から言えばアルゴリズムの盗用が大きな問題となるため，これらの手法が有用であることは疑いようもない．しかし，プログラミングの授業では同一の目的でプログラムを作成するため，自ずとアルゴリズムが類似してしまう．そもそも，同一のアルゴリズムをプログラミングする場合も多い．また，授業で作成するプログラムはアルゴリズムが特定できないほど短いものも多い．つ まり，授業で作成されたソースコードに対しては，既存のソースコードの盗用発見手法を適用することはできない．

申請者らはこれまで機械学習に関する研究を進めてきた．その中で，応用の１つとして，参照データを用いたソースコードの類似性発見手法を提案して来た．これは上述の分類で言えば，1）の文字列マッチングに基づく手法である．しかし，これを，採点支援を目的として，プログラミング授業における提出課題に対して適用した際には，前記のような理由により十分な性能が発揮されなかった．この解決策として，従来法では有効に利用されてこなかった表面上の特徴に着目するという考えを思いついた．これらの特徴を抽出するために，強化学習や動的プログラミングの研究で用いた確率過程モデルが利用できると考えた．

２．研究の目的

本研究では，まずプログラミング授業の課題に適したソースコード盗用発見アルゴリズムの構築を行う．すなわち，アルゴリズムの類似性が盗用の根拠とならない場合や，ソースコードが短いためにアルゴリズム上の類似性を発見することが困難な場合にも利用できるような盗用発見手法の開発を行う．

本研究ではこの目的のために，インデントや空白，改行の使用法やコメントの形式など，ソースコード作成者の記述時の“くせ”とも言えるような「表面的な特徴」を利用して，作成者の記述に関する確率モデルを作成する．疑わしいソースコードが，実際にその作成者によって書かれたものか，それとも盗用であるかは，このモデルへの合致度合いに基づいて判定する．

続いて，このアルゴリズムを用いた，プログラミング授業におけるソースコード盗用の可能性を提示する採点支援システムを試作し，このシステムを申請者が担当している実際のプログラミング授業における課題プログラムを用いて評価する．

３．研究の方法

３．１ アルゴリズムの構成

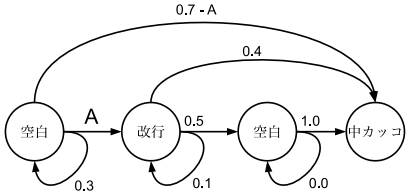
本研究の提案する手法では，あらかじめ作成者が明らかな複数のソースコードを用いて，その「表面的な特徴」から作成者の「記述モデル」を作成する．そして，新たなソースコードが得られる度に，そのソースコードの作成者とされる人物の記述モデルと照らし合わせ，ソースコードが実際に記述モデルの対象となっている作成者によって書かれたものであるかどうかを判定する．

記述モデルでは，特定の記号の前後に空白を幾つ入れるか，といった記述上の特徴をモデル化する．この目的では，幾つかの確率モデルを利用することができるが，本研究では，まず，最も単純な，隠れマルコフモデルを利用する．隠れマルコフモデルは，記号列の出現確率をモデル化する確率モデルであり，Baum-Welch アルゴリズムを用いることにより，記号列の出現確率を実例から学習することができる．

例えば，下図に示した非常に単純な隠れマルコフモデルの例は，「1 個以上の空白→0 個以上の改行→0 個以上の空白→中カッコ ({)」という記号列の出現確率をモデル化する．もし，あるソースコードの作成者が「中カッコの前には必ず改行を挿入する」という記述上

の特徴を持っているとすれば，図中の「A」の遷移確率が高くなる．

このように，隠れマルコフモデルを用いることで，ソースコードの「表面上の特徴」をモデル化することができると考えられる．



隠れマルコフモデルの構造が決まれば，モデル化される記号列，すなわち，ソースコードの「表面的な特徴」として利用される記述上の特徴もある程度固定される．したがって，本研究において，隠れマルコフモデルの構造は非常に重要である．

本研究では，オープンソース・プロジェクトのソースコード，および，従来研究で用いたプログラミング授業の課題ソースコードを対象として，記述上の特徴を抽出し，これを元に隠れマルコフモデルの構造を決定した．

３．２ 実装とテスト

定式化したアルゴリズムをワークステーション上に実装し，これを用いて，実装したアルゴリズムがソースコード作成者を識別できるかテストを行った．

すなわち，申請者の研究室の学生を対象とした演習において，学生に幾つかの課題を与え，ソースコードを作成させる．このうち，各学生について，提出されたソースコードの幾つかを用いて記述モデルの学習を行い，未学習のソースコードを用いてテストを行った．

４．研究成果

例えば，特定の記号の前後に空白を幾つ入れるか，といった記述上の特徴をモデル化する目的で，本研究では隠れマルコフモデルを利用したが，その構造は，従来研究で用いたプログラミング授業の課題ソースコードから記述上の特徴を抽出し，これを元に決定した．また，未知の特徴が現れるたびに新たな記述モデルを追加するという適合学習を導入した．

提案手法に基づいたソースコード作成者認識システムを実装し，研究室の学生のプログラムを用いてテストを行った結果，平均して 80%以上という認識率が得られた．

本手法では，従来，ソースコードの比較によってしか行えなかった盗用発見の試みを，単一のソースコードに対して行える．また，記述上の特徴を用いるため，プログラミングの授業などで作成される，短く，類似した構造のソースコードに対して利用できる．これにより，授業の円滑な進行と，提出課題の適正な評価を支援することができる．

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 10 件）

① Yancong Su, Hajime Murao, A Study on Human Behavior Recognition using Discrete HMM and Continuous HMM, ICIC Express Letters, 査読有, 6 巻 3 号, 2012, 791-796

② Masato Nagayoshi, Hajime Murao, Hisashi Tamaki, Adaptive Co-construction of State and Action Spaces in Reinforcement Learning, Proc. of the 16th Int. Symp. on Artificial Life and Robotics, 査読有, 2011, 589-592

③ Asako Ohno, Hajime Murao, A two-step in-class source code plagiarism detection method utilizing im- proved cm algorithm and sim, Int. J. of ICIC, 査読有, 7 巻 8 号, 2011, 4729-4739

④ 永吉雅人, 村尾 元, 玉置 久, 連続行動空間への適用を考慮した Switching 強化学習, 電気学会論文誌 C, 査読有, 131 巻 5 号, 2011, 976-982

⑤ Asako Ohno, Tsutomu Inamoto, Hajime Murao, Fundamental study of clustering images generated from customer trajectory by using self-organizing maps, ICIC Express Letters, 査読有, 4 巻 5B 号,2010, 1979-1984

⑥ Yancong Su, Hajime Murao, Estimating energy expenditure of mobile device users using HMM, roc. of the 2nd Int. Conf. on Intelligent Computing and Intelligent Systems, 査読有,2010,CD-Paper

⑦ Asako Ohno, Hajime Murao, An author identification of in-class source codes by using the forward- backward coding models, Proc. of the 5th Int. Conf. on ICIC, 査読有, 2010, CD-Paper

⑧ Asako Ohno, Hajime Murao, Work in progress - a novel methodology to reduce instructors' and students' psychological burdens in source code plagiarism detection, Proc. of the 40th Annual Frontiers in Education Conference, 査読有, 2010, S3D1-S3D2

⑨ Asako Ohno, Hajime Murao, A new similarity measure for in-class source code plagiarism

detection, Int. J. of ICIC, 査読有,5 巻 11B 号, 2009, 4237-4247
⑩ Asako Ohno, Hajime Murao, Modeling and quantification of superficial features extracted from source codes: In consideration of fluctuation of description among learning data, Proc. of the 4th Int. Conf. on ICIC, 2009, CD-Paper

〔学会発表〕(計１件)

① 蘇 彦聡, 筒井智子, 村尾 元, 携帯端末ユーザのエネルギー消費量推定への HMM の適用, 第 37 回知能システムシンポジウム, 2010 年 3 月 16 日, 横浜/日本

６．研究組織

(1)研究代表者

村尾　元（MURAO HAJIME）
神戸大学・大学院国際文化学研究科・准教授
研究者番号：70273761

(2)研究分担者

該当無し

(3)連携研究者

該当無し