様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２８ 年　 ６ 月 ２７ 日現在

機関番号：１４５０１
研究種目：基盤研究(C)（一般）
研究期間：2013～2015
課題番号：２５３３０４０８

研究課題名（和文）記述モデルに基づいた レポート盗用発見シス テムの構築

研究課題名（英文）Development of a Plagiarisms Detecting System for Reporting Assignments based on the Style Model

研究代表者
村尾　元（Murao, Hajime）

神戸大学・国際文化学研究科・教授

研究者番号：７０２７３７６１

交付決定額（研究期間全体）：（直接経費）　　3,600,000円

研究成果の概要（和文）：本研究では，授業における課題レポート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レポートから，句読点の利用法など，レポート作成者の記述上の特徴を，隠れマルコフモデルを用いて学習する。そして，新しいレポートが得られる度に，その作成者とされる人物の記述モデルと照らし合わせ，レポート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テストでは70%程度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 plagiarisms in reporting assignments. The system firstly train a Hidden Markov Model (HMM) for each of students from "expression feature" of submitted reports, such as a number of punctuations, the location of new-lines, etc. We call the trained HMM a "style model". Then, whenever a student submits a new report, the system uses the student's style model to validate if the report has been actually written by the student. We have tested the system in our laboratory and observed quite fair result around 70% of correct classification.

研究分野：社会システム科学

キーワード：盗用発見　機械学習　教育支援

１．研究開始当初の背景

コンピュータとインターネットの発展とともに，学校のレポート課題におけるコピーすなわち盗用が問題となってきている。今日では，与えられた課題に関連するいくつかのキーワードを指定して検索するだけで，インターネット上に散在する，様々な情報や他人の主張が書かれた文章を，手元のコンピュータに取り寄せることが可能である。このように検索された文章はマウスとキーボードの簡単な操作でコピーして自分の文書に貼り付けることができる。

授業の効果を上げるためには，提出されたレポートの適正な評価が欠かせないが，大量のレポートを対象とした盗用発見の試みは教員への負担が大きい。したがって，適正な評価の実施と，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は必要不可欠である。

一方，社会的，経済的な影響も大きいことから，文書の盗用発見へのコンピュータの適用については，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) Web や他の文書などの外部コンテンツとの一致検出に基づく方法と 2) 同一文書内における一貫性の有無に基づく方法に大別できるが，そのいずれもが，単語や文を比較し，その一致や独自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，文章とその伝える内容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，授業のレポート課題では，同一または限られた少数のテーマに基づいてレポートを作成するため，使用する単語や文は類似してしまう場合も多い。そもそも，利用する単語や文がある程度指定されている場合すらある。また，授業で作成するレポートは，単語や文の比較からは出典を特定できないほど短いものも多い。つまり，授業で作成されたレポートに対して，既存の文書盗用発見手法の適用は困難である。

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，プログラミング授業におけるソースコードの盗用発見手法を提案してきた。提案手法では，インデントや空白，改行の使い方や，コメントの形式といった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し，これに確率モデルの考え方を適用した。その過程で，表面上の特徴に着目するという我々の提案手法が，プログラミング言語という人工言語のみならず，自然言語で書かれた文書の盗用発見にも利用できるのではないかと考えた。

２．研究の目的

本研究の目的は，レポートにおける盗用発見システムの構築である。すなわち，課題として学生から提出されたレポートを対象として，盗用の可能性を推定し，これを教員に提示するシステムの試作である。これにより，教員の負担を軽減しつつ，適正な採点および授業の実施を支援する。

本研究の中心的な課題は，授業のレポート課題のように，単一もしくは非常に限定されたテーマに基づいて書かれた自然言語の文書から盗用の可能性を推定する手法の開発である。このような文章では，その内容の類似性が盗用の根拠とならず，また，レポートが短いために文章やその内容上の類似性を発見することが困難となる。このよう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，句読点や空白，改行の使用法といった，レポート作成者の記述時の“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。疑わしいレポートが，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。

また，このアルゴリズムの実用性を検討するため，授業における課題レポートの盗用可能性を提示するシステムを試作し，このシステムを実際的な環境で評価する。

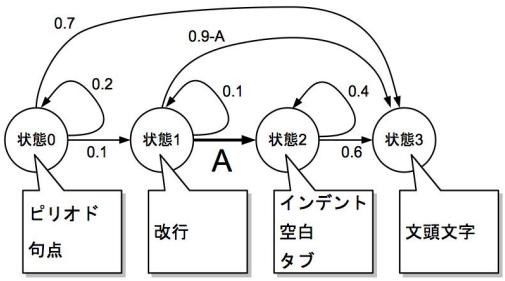
３．研究の方法

３．１．アルゴリズムの概要

本研究の提案する手法では，まず，レポートの「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，新たなレポートが得られる度に，そのレポートの作成者とされる人物の記述モデルと照らし合わせ，レポート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，例えば，特定の位置に空白を幾つ入れるか，句読点の比率はどれくらいか，改行の頻度はどれくらいか，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，従来研究でも利用してきた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 個以上の句点→0 個以上の改行→0 個以上の空白→文頭」という記号列の出現確率をモデル化する。もし，あるレポートの作成者が「文頭には必ず空白を挿入する」という記述上の特徴を持っているならば，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，レポートの「表面上の特徴」をモデル化する。

３．２．アルゴリズムの構築と検討

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，レポートの「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，まずは，インターネット上の文章，例えば Wikipedia や著名な解説サイトの文章，および，これまでの授業で得られた課題レポートを対象として，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。

続いて，定式化したアルゴリズムをパソコン上に実装する。実装に関しては，特に外部のパッケージ等を利用せず，申請者自身がプログラミングを行った。その後，申請者の研究室の学生を対象として，実装したアルゴリズムがレポート作成者を識別できるかどうかテストを行った。すなわち，学生に幾つかの課題を与え，レポートを作成させる。このうち，各学生ごとに，提出されたレポートを用いて記述モデルの学習を行い，未学習のレポートを用いてテストを行う。このテストにおいて，アルゴリズム全般，特に隠れマルコフモデルの構造について検討を行った。

３．３．盗用発見システムの構築とテスト

盗用発見システムを Web アプリケーションとして構築した。学生は特定の Web サイトに自分のユーザ ID（主に学籍番号で構成される）でログインし，指示に従って課題レポートを転送する。これらは Web サーバ上のデータベースに登録される。教員は同じ Web サイトに自分のユーザ ID でログインし，提出されたそれぞれのレポートについて，盗用の可能性に関する情報が確認することができる。すなわち，学生 A が提出した課題が，どの学生の記述モデルと合致するか，学生 A の記述モデルとはどの程度合致するのかが提示される。

複数の学生を対象に，教員役，学生役を割り当てながら，テストを行い，ユーザインターフェイスや運用法の検討を行った。

４．研究成果

本研究では，授業課題のレポートのように，内容による比較が困難な文章に対して，「表面上の特徴」による盗用発見手法の提案を行った。すなわち，レポートの盗用発見における，句読点や改行，空白などの使い方といった書き方の“くせ”の利用を試みた。

この目的で隠れマルコフモデルを利用した。すなわち，提出されたレポートにおける句読点や改行，空白などの出現確率を隠れマルコフモデルにより学習するとともに，これを提出されたレポートへの合致度を調べることで，レポートがモデル化されている作成者に書かれたものかどうかを判定する。

提案手法に基づいたレポート作成者認識システムを実装し，研究室の学生を対象にテストを行った結果，70%程度という認識率が得られた.

本手法では，従来，比較によってしか行えなかった盗用発見の試みを，単一のレポートに対して行える。また，記述上の特徴を用いるため，授業などで作成される，短く，内容的に類似したレポート対して利用できる。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５件）

(1) Hajime Murao. Evaluating city personality from geo-tagged sns messages. ICIC Express Letters, 9(3):729-735, 3 2015.

(2) Hajime Murao. Personality estimation from sns messages and its application to evaluating a city personality. Procedia Technology, 18:72-79, 12 2014.

(3) Masato Nagayoshi, Hajime Murao, and Hisashi Tamaki. An entropy-guided adaptive co-construction method of state and action spaces in reinforcement learning. volume 8834 of Lecture Notes in Computer Science, pages 119-126. Springer, 2014.

(4) D. Moritz Marutschke and Hajime Murao. Short study on complexity and feasibility of deterministic epidemiological models to track knowledge propagation in scientific publications. ICIC Express Letters, 8(4):1081-1088, 4 2014.

(5) Yancong Su and Hajime Murao.

Extracting feature values from human gait by using lpc cepstrum analysis for attribute recognition. ICIC Express Letters, 8(3):815-820, 3 2014.

〔学会発表〕（計１件）

(1) Shuhei Miyake and Hajime Murao. Hedgehog: Team formation system learning performance of team. Proc. of International Conference on Innovative Computing, Information and Control (ICICIC2015)

〔図書〕（計０件）

〔産業財産権〕
○出願状況（計０件）
○取得状況（計０件）

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・教授
研究者番号：70273761

(2)研究分担者
なし

(3)連携研究者
なし