授業レポート提出システムへの追加を想定した盗用検出機能の開発

Development of a plagiarism detection function for the online submission system

テーマ：インターネット技術とその応用　　**教養学部　情報科学科**
指導教員：松本 章代　　1257237　**鈴木 慶太**

1. 研究背景

現代の大学教育の現場において，学生がレポート課題に取り組む際に行われる盗用が問題となっている．また，課題の採点作業において，盗用の確認作業が教員の大きな負担となっている．

そこで本研究では，既に本研究室で開発したレポート提出システムへの追加を想定した盗用検出機能の開発を行う．これによって学生が盗用を行うことに対しての抑止力として働かせることが可能になる．また，教員が採点を行う際の負担を軽減することができる．

従来より，レポートの盗用を検出するシステムは，数多く存在している。これらの中には，インターネット上のテキストからの盗用も検出する機能が備わっている製品もある．しかし，これらの製品は学生が提出したレポートを改めてソフトウェアで検出作業を行う必要がある．これらの既存ツールとの相違点として本研究では，レポート提出システムと連携させることで，提出されたレポートをそのまま盗用の検出を行うことを想定している．これによって，教員の負担をより大きく減らすことができる．

2. 関連研究

一般的に文書間の類似度を求める手法として，TF-IDF，ジャッカード係数，シンプソン係数，ダイス係数，トライグラムなどが用いられている．

TF-IDF とは，文書中の単語に関する重みの指標である．TF と IDF の積で求められる．TF とは，文書の中でその単語がどれだけ多く登場するかを表す指標である．IDF とは，その単語がどれだけ特徴的かを表す指標である．例えば,「する」や「ある」などは多くの文章に含まれるため重要度が低くなり，逆に一つの文書にしか登場しない単語は重要度が高くなる．この二つの指標を掛け合わせて求められるのが TF-IDF である．

ジャッカード係数，シンプソン係数，ダイス係数の 3 つは，文書を単語で区切り集合として扱い，和集合や積集合を求めて計算を行うことで，2 つの文書の類似度を求める指標である．

トライグラムとは，文書を単語単位ではなく 3 文字単位で分解し，出現頻度を求める手法である．単語や文節での分解とは異なり，文字単位で取り出すため，プログラム上での実行時間が短くなる．

太田ら [1] は,「模倣レポート判定に用いる文書間類似度」を考案した．最小構成文を生成し, 最小構成文間類似度を求めるという独自の手法を提案している．

本研究では，類似度を求める手法だけでなく，TF-IDF を用いて特徴的なフレーズを抽出する手法を採用している．人間が実際に盗用の確認を行う際には，誤字脱字や，文体の間違いなど，通常は見れらないような特徴的なフレーズが複数のレポートで一致する場合に怪しいと考えることが多い．特徴的なフレーズを抽出することによって，人間が行う場合に近い手法で盗用検出を行うことができる．この手法は関連研究では行われていない．

3. システム概要

本システムは，授業レポート提出システムによって提出されたレポートに対し，盗用の判定を行う．その際，少し怪しいと思われるレポートと，ほぼ間違いなく盗用であると思われるレポートの二段階の評価を行う．盗用検出機能の開発には，他のプログラミング言語に比べてテキスト処理に適した Ruby を用いた．また，レポートの内容を単語や文節に区切って行う分析には，一般的には文の係り受けの解析に用いられている CaboCha を使用する．

4. 盗用検出の手法

盗用検出に用いる手法として，TF-IDF，トライグラム，シンプソン係数の 3 つの指標を採用した．それぞれの式は表 1 に表す．

4.1 TF-IDF

TF-IDF は本来，単語の重みであるが，本研究では文節毎の重みを求めた．レポート課題において，教員がテーマを指定した場合には使われる単語が似ることがある．その場合に単語に重み付けを行ってもあまり差が出なくなってしまう．一方，文節で区切った場合には，その単語に続く接続詞の使い方や，言い回しに特徴が表れる．そのため単語ではなく文節での重み付けを行った．こうして求めた TF-IDF が高い文節を特徴的なフレーズとし，抽出する．同じフレーズが抽出された文書が盗用の疑いがあると考える．

4.2 シンプソン係数

ジャッカード係数，シンプソン係数，ダイス係数の 3 つは，どれも文書を単語単位に区切って類似度を求める手法だが，計算式が少しずつ異なるため，値の出方に差が出る．二つの文書の単語数が極端に差があると，類似度が低いものでも高く検出されてしまう傾向があるが，その影響が大きいのがシンプソン係数である．

4.3 トライグラム

トライグラムは単語や文節の区切りを無視するため，「京都へ旅行に行った」という文の「京都へ」と「東京都へ旅行に行った」という文の「京都へ」が同じと見なされてしまうなどの欠点がある．しかし，単語で区切った場合とは異なる視点で見ることで，盗用を見逃

す可能性を下げることができると考える．

表 1.　計算式

指標	式
TF	単語数/出現回数
IDF	$-log$(ある単語が含まれる文書数/全文書数)
TF-IDF	$TF \times IDF$
トライグラム	$X \wedge Y / X \vee Y$
シンプソン係数	$X \wedge Y / min(X, Y)$

4.4　盗用かどうかの境界値

上記の 3 つの指標で求められた数値は，扱う文書の文字数や文書数によって変化する．したがって盗用と判定する境界値を常に一定にすることはできない．そこで，以下のような手法を取る．

1. 求められた値を降順にソートする．
2. ソートした値の隣同士の値の差を求める．
3. 差が最も大きいところを分割区間とする．
4. 差が小さい場合に盗用なしとするための定数 n を定める．
5. 求めた分割区間が n 以上か．
   - n 未満なら盗用なし．
   - n 以上なら 6. へ．
6. 分割区間より値が小さい方と大きい方のどちらがデータが多いか．
   - 小さい方が多い場合は大きい方を盗用とみなす．
   - 大きい方が多い場合は大きい方の集合のみを対象として 3. へ戻る．

4.5　提案する盗用検出手法

過去の学生のレポートを用いて事前に実験を行った．2012 年度の初年次教育でのレポート課題を用いて行った．文書数は 94 である．その結果を表 2 に示す．

表 2.　実験結果

指標	HIT	FA	MISS	CR	精度	再現率
TF-IDF	10	10	0	74	66.7%	100.0%
トライグラム	6	0	4	84	100.0%	60.0%
シンプソン係数	5	0	5	84	100.0%	50.0%

トライグラムとシンプソン係数がどちらも精度が高く，再現率は少し低くなるという結果が出た．一方 TF-IDF は精度は少し低くなるが，再現率が 100.0%となっている．

この結果から，オリジナルの手法では，再現率が高い TF-IDF で検出された文書を盗用の疑いがあるとする．また，精度の高いトライグラムとシンプソン係数の二つで検出された文書を，それぞれを集合と考え，和集合を求める．その結果を盗用の可能性が高いと判定する．これによって，盗用を見逃す可能性も低くしつつ，より盗用の可能性が高いものに関しては強調することができると考える．

5.　評価実験

実験で未使用であったレポートを用いて，オリジナルのプログラムの評価実験を行う．

5.1　実験に用いるレポート

表 3 は評価実験で使用したレポートである．どちらも 2013 年度のプログラミング中級のレポート課題で，A の文書数は 79 で，B の文書数は 84 である．

表 3.　評価実験で使用したレポート

記号	授業名	授業回
A	2013 年度プログラミング中級	第 6 回
B	2013 年度プログラミング中級	第 14 回

5.2　実験結果および考察

評価実験の結果を表 4 に示す．

表 4.　評価実験

授業回	盗用の可能性	HIT	FA	MISS	CR	精度	再現率
A	疑い	3	6	0	70	33.3%	100.0%
A	可能性大	3	3	0	73	50.0%	100.0%
B	疑い	0	13	10	61	0.0%	0.0%
B	可能性大	10	0	0	74	100.0%	100.0%

盗用の可能性が高いという判定に関しては，いずれも再現率で 100.0%を達成することができた．精度においても，B のレポートでは 100.0%を達成し，A のレポートでも 50.0%と，ある程度の成果を上げることができた．これは，文字単位と単語単位という異なる視点から見ることで，精度の高さを維持しつつも，盗用の可能性の高いものを見逃さないようにしたいという思惑通りの結果となった．しかし，盗用の疑いがあるという判定においては，B において精度，再現率ともに 0.0%という結果が出てしまった．これは，TF-IDF を用いたプログラムが特徴的なフレーズという一点にのみ着目して判定を行っていることが原因だと思われる．特徴的なフレーズが一致していれば，それ以外がかけ離れた内容であっても盗用であると判定してしまい，逆に特徴的なフレーズがなければ，内容が似ていても盗用ではないと判定されてしまう．

TF-IDF を用いたプログラムを採用した理由として，取りこぼしを少なくしたいという考えがあったため，再現率が低くなってしまったのは当初の想定とは異なる結果になってしまった．

6.　まとめ

本研究では，文書間の類似度を求める様々な手法について検証を繰り返し，その中から 3 つの指標を用いて，盗用検出機能を開発することができた．そして評価実験で一定の成果を上げることができた．しかし，盗用の疑いの検出において再現率が低くなるという結果が出てしまった．したがって，今後この研究を進めるとすれば，盗用検出の手法に関して，さらに調査を行う必要があるだろう．また，既存のレポート提出システムへの追加も完了していないため，これも進めていく必要がある．

参考文献

[1] 太田 貫久，増山 繁：模倣レポート判定に用いる文書間類似度の考案，言語処理学会年次大会発表論文集，Vol.10th，No.CD-ROM，pp.A10B6-03 (2004).