E-10 学生レポートの n-gram による類似度評価の検討

村田 哲也† 黒岩 丈介‡ 高橋 勇‡ 白井 治彦‡ 小高 知宏‡ 小倉 久和‡
T. Murata, J. Kuroiwa, I. Takahashi, H. Shirai, T. Odaka, H. Ogura

1. まえがき

学生の中には、他人と全く同じレポートや、一部だけ書き換えてレポートを提出する人がいる。そこで、テキスト解析の一手法である n-gram 解析を用い、これらのレポートがどれくらい他の人が書いたものと類似しているかを評価する方法を検討する。そして、実際に評価関数を用いて類似度を求めることにより、効率的に他の人が書いたものを写したかどうかを発見できることが分かった。

2. 文章の解析方法

レポート間の類似度を評価する方法として、 n-gram 解析を用いた。 n-gram 分布とは n 個の文字が隣接して生じる文字の共起関係、すなわち n-gram の出現頻度を記録したものである。その n-gram 分布を元に後述の評価関数を用いて、評価関数を計算し、得られた評価値を文章間の類似度とする。そして、得られた類似度から形式的にどの程度、類似しているかを調べる。

文章全体は、 n 文字からなる文字パターン $Xi(i = 1, 2, \cdots N)$ の連鎖で表すことができる。つまり文章が N 個の文字パターンの連鎖から成るとすると文章全体 x は、

$$x = \{X_1, X_2, \cdot\cdot, X_i, \cdot\cdot, X_N\}$$

である。これを用いて、文章 A,B 間の類似度を以下のように定義する。

$$R = 1 - \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{P_A(X_k) - P_B(X_k)}{P_A(X_k) + P_B(X_k)}\right\}^2 \qquad (1)$$

但し、 $N = max(N_A, N_B)$, $P_A(X_k)$ 及び $P_B(X_k)$ は、各々、文章 A 及び文章 B 中の文字パターン X_k の出現頻度分布である。この類似度定義では、文章 A と文章 B の出現頻度分布が全く同じである時、 R = 1 となる。逆に文章 A と文章 B の n-gram 分布に現れる文字のパターンが一つも重複しない時、 R = 0 となる。 今回は、 n=3 の 3-gram 解析によって、類似度の評価を行なう。レポートの文字数は、約 400 文字であり、日本語の場合、機能する実質語の多くは、 3 文字で構成されているので 3 が妥当だろうと考えた。

3. n-gram による類似度評価

3.1 シミュレーション

まず、学生が行うレポートの書き換えに関して、検討を行い、その書き換えを機械的に行うフィルタを作成した。そして、変換前と変換後のレポートを比較し、本システムがこれらの書き換えに対して正しく類似度を与えるか確かめた。実際に行ったフィルタ操作は、

- 文末の変換
- 単語の置換
- 文章の出現順序の入れ替え
- 文章の挿入
- 文章の置換

である。以下に上述のフィルタ操作を行った文章に対する類似度を評価した結果の特徴的なものを数例与える。ここで文章の変換率とは、レポートの文章全体に対する変換した文字の百分率であり、以下のように定義される。

$$\text{文章の変換率} = \frac{\text{変換した文字数}}{\text{レポートの文章全体の文字数}}$$

3.1.1 文末変換に対する類似度

ファイル名リストからレポートファイルを取り出し、レポートを区点毎に読み込む。次に事前に用意した文末のパターンと読み込んだ文末がマッチすれば文末を変換する。図 1 にオリジナルの文章を文末変換を施したもの

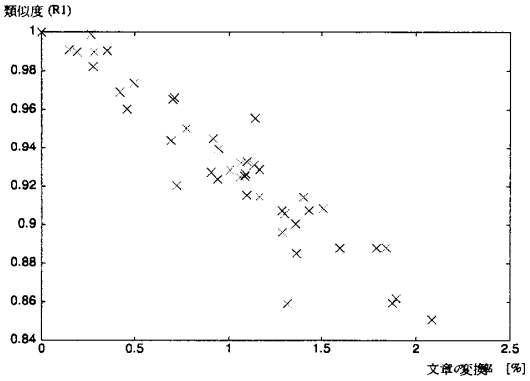

図 1: 文末変換文の類似度

のとの類似度をを与える。全ての場合に対し、類似度が 0.84 程度と非常に類似度が高く、文末を書き換えたレポートは、本手法により発見できることが期待される。

3.1.2 単語の置換に対する類似度

コーパスフィルタ「茶筅」を用いて、各レポート文章を品詞分解する。そして、名詞及び未知語の中で最も出現頻度の高いものを全く意味の無い文字列 (例えば「AAA」) に置き換えた。この場合、オリジナルの文章との類似度を図 2 に与える。図より全ての場合に対し類似度が 0.76 程度と高いことが分かった。図 2 より、頻出する単語の書き換えを行っても、類似度は、0.76 以上となった。

†福井大学大学院 工学研究科 情報工学専攻
‡福井大学 工学部 知能システム工学科

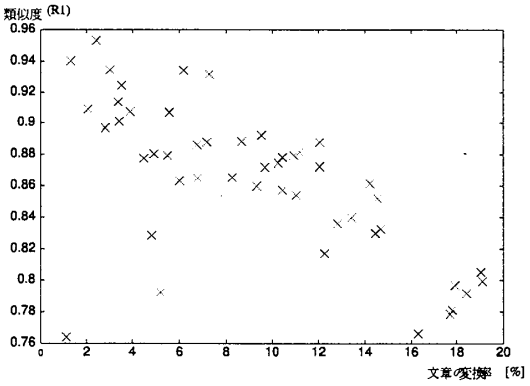

図 2: 単語の置換文の類似度

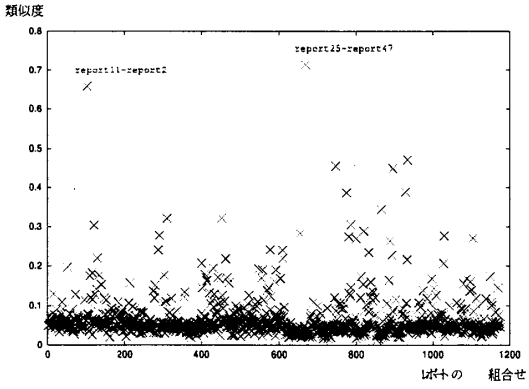

図 4: 学生レポート実験の結果

3.1.3 文章の出現順序の入れ換えに対する類似度

レポート文章中を一文毎に分離し、それらっをランダムに入れ換えて、新たな文章を作成した。この場合の類似度 (図 3) も 0.9 程度と非常に高い値であった。次に文章を乱数を用いて、ランダムに入れ換える。図 3 の結

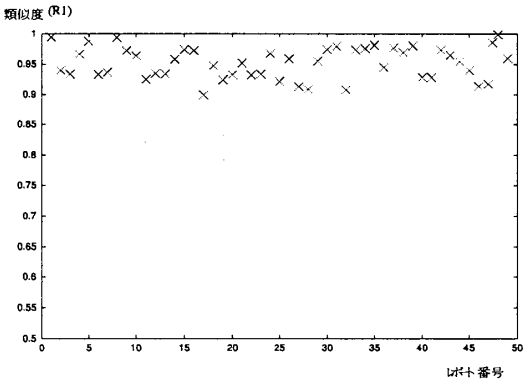

図 3: 文章の出現順序の入れ換え文の類似度

果から分かるように、元のレポートとレポートの文章の出現順序をランダムに入れ替えてたレポートとの類似度は 0.9 以上になった。

3.2 学生レポートに対する実験

最後に実際の学生レポートをつかって実験した。使用したレポートは知能工学科 1 年生の後期の計算機システム講義に出題されたものである。結果を図 4 に示す。図より、 report11 と report2 、 report25 と report47 の組合せが極端に類似度が高くなった。実際にこれらの組合せを詳しく調べてみると、類似度とした。この結果の類似度の高いレポートの組合せは、文末が書き換えてあったり、1、2 語付け加えてあったり、単語を違う語に変換したりしたものであった。

4. 考察とまとめ

本研究では、教科書のコピーや他人のレポートのコピー、インターネットの web サイトのコピーがレポートの不正行為であると考える。その不正行為の方法として文末の変換、単語の置換、文章の順序の入れ替え、もしくは文章の挿入や置換が主なものであると考えた。実際、学生レポートに対して本手法を適用することにより、高い類似度がでることが分かった。つまり、考えられる学生の不正レポートに対して、類似度比較システムで計測を行なえば、類似度が比較的高い結果が得られることが分かった。また、実際に学生が提出したレポートを用いてこのシステムを試したところ、高い類似度の組合せが発見できた。よって、実際の現場での学生レポートに対しても n-gram 解析を用いたレポート解析が有効であることが示された。

今後は、まず n-gram の n の値を変えた際に類似度にどのような特徴が取れるか研究を行う予定である。更に、n-gram に変わる文章の解析方法を考え、類似度評価方法の検討をしていこうと考えている。

参考文献

[1] 福田泰弘 , 西野順次, 小高知宏, 小倉久和
「テキストの類似度比較の定義とその自動抽出の試み」 平成 12 年度卒業研究,2000.

[2] 松浦 司, 金田 康正
「近代日本小説家 8 人による文章の n-gram 分布を用いた著者判別」 (情報処理学会研究報告, pp1-8,2000)