白梅学園大学・短期大学情報教育研究
2019, No. 22, 25－30.

WebClass類似レポート検知機能の検証

倉澤　寿之

本学のLMS（Learning Management System）として2018年度から採用されているWebClassには、提出されたレポートの内容がどの程度類似しているのか判定する機能（「類似レポート検知」）が備わっている。大学の授業課題としてのレポートを手書きではなく、パソコンやスマホで作成するのがふつうとなり、ネットからの引用も多くなった昨今においては、剽窃のチェックに必要な機能だと言える。本稿では、類似度を加工した模擬レポートを作成してこの機能によるチェックを行い、判定の有効性を検証することとした。

使用した模擬レポート

チェックの対象とした模擬レポートは、筆者が「発達臨床実験法」という授業で学生に提示している心理学実験レポートの記述例２種（「ミューラー・リヤーの錯視」と「ストループ効果」）である。この授業は心理学の方法論としての実験法を実際に自分たちのデータを取って学ぶもので、これら２種のテーマは最初の段階の入門編で取り上げられている。学生にとって、「目的」「方法」「結果」「考察」「要約」「引用文献」という構成が決められた心理学の実験レポートを書くのは初めての体験であるため、入門編では典型的なデータを元に記述したレポート記述例を提示し、最初の段階ではこの記述例に沿ってレポートをまとめるように指示している。その記述例として使用しているレポート２種を今回のチェックに用いる主なレポートとした。これ以後、「ミューラー・リヤーの錯視」レポート記述例を「MLレポート」または単に「ML」、「ストループ効果」レポート記述例を「STレポート」または単に「ST」と略記することにする。レポートタイトル、氏名、学籍番号などの書かれた表紙を別にすると、それぞれの文字数はMLレポートが1976文字、STレポートが3231文字である。以降の類似レポート検知機能のチェックでは、これらの模擬レポートに人為的な変更を加えたものを使用した。

他方、これら２種の模擬レポートは心理学実験という同一の分野に属するものであり、使われている用語（「被験者」「手続き」「分散分析」など）にかなり共通性がある。そこで、これらとは全く別分野となるレポート、具体的には筆者が「情報教育研究」に発表した「剽窃検出ソフトWCopyfindの日本語文書への適用について」（倉澤,2016）を別分野のレポートとして取り上げ、分野の全く異なるレポート間での類似度チェックに使用した。このレポートを以下「別分野レポート」と呼ぶ。

なお、WebClassの類似レポート検知機能はpdf形式の文書やテキスト形式も扱えるが、今回の検証では、模擬レポートをいずれもWordの文書ファイル（.docx）の状態でWebClassにアップロードした。

類似度の操作とその結果

以下、類似度を変更するために行った操作と、その結果としての類似度スコアを紹介する。なお、WebClassでは類似度スコアと剽窃の可能性について、表１のような目安を挙げていてる。

表１ WebClassの類似度スコアの目安

スコア	剽窃の可能性
85～100	強い剽窃の疑いがあります
70～ 84	剽窃の疑いがあります
40～ 69	引用など近似した段落が含まれます
0～ 39	剽窃の可能性は低いです

①変更前の元のレポート間の類似度

類似度を人為的に変更する前の状態でのスコアを確認しておくため、使用した３種のレポート間で類似度チェックを行ったところ、類似度スコアはMLとST間で24.0、別分野レポートとの間ではMLが11.6、STが13.6となった。図１にMLとST間での「差分表示」を示す。この差分表示を見る限り、類似度は文字単位の重複度合いが元になっているものと思われる。個々の文字単位での重複度合いとは別に、文字の連続性、つまり単語レベルでの重複が評価対象になっているかどうかはわからないが、同じような用語が使用される同一分野のレポート間のほうがスコアは高くなっている。

②完全に同内容のレポート間の類似性

MLレポートとSTレポートのそれぞれについて、ファイル単位でコピーした全く同一内容のファイルを２つ作り、別々の学生アカウントで提出して類似度チェックを行ったところ、類似度スコアはMLレポートで85.3、STレポートで82.0となった。

この結果は意外である。ファイルコピーした完全に同じ内容のレポートなので、類似度スコアとしては100を期待したいところだが、実際にはMLで「強い剽窃の疑い」の範囲にかろうじて入ったものの、STでは「剽窃の疑い」という一段下の評価段階にとどまるスコアであった。


» 結果一覧に戻る

類似レポート検知 結果

内容1	内容2
「ミューラー・リヤーの錯視」白梅学園大学子ども学部発達臨床学科Z0Z001 白梅香目的　ミューラー・リヤーの錯視(Müller-Lyer,1889)図形 (図1)において、矢羽の角度と長さが主線の長さの見え方に及ぼす効果を検討することを目的とする実験を行った。(主線a)(主線b)(主線a)(主線b)外向図形外向図形内向図形内向図形1 ミューラー・リヤーの錯視図形図1 ミューラー・リヤーの錯視図形方法実験日時 20xx年x月xx日 午後x:xx～x:xx被験者 大学生18名器材 ミューラー・リヤー錯視測定図形((株)竹井機器製)。図版は、矢羽と主線の夾角が30度で矢羽の長さが15mm(A)、35mm(B)、45mm(C)の3図形と、矢羽の長さが30mmで、夾角が15度(D)、30度(E)、60度(F)の3図形の6図形(表1)からなり、主線の長さはいずれも100mmであった。表1 実験に用いたミューラー・リヤーの錯視図形図形ABCDEF矢羽の長さ(mm)153545 30夾角(度)30 15 30 60手続き 調整法を用いて実験を行った。内向図形(図1の(a))を標準刺激、外向図形(図1の(b))を比較刺激とし、標準刺激を左に、比較刺激を右にする空間順位で行った。各図形につき、上昇系列、下降系列、下降系列、上昇系列の4系列を連続して行った。6枚の図形の呈示順序は、被験者によってランダムに変えた。次のような教示を与えて実験を行った。『これから実験を始めます。矢羽で区切られた線の右側と左側の長さが等しく見えるように、右の部分の長さを長くしていったり、短くしていったりして調節してください。長さが等しく見えたところで止めて、図版をこちらに渡してください。』錯視量は、外向図形の主線が内向図形よりも短い場合を負の値として、ミリメートル単位で記録した。結果　まず、各被験者の測定値について4系列のデータを平均し、図形ごとの錯視量を算出した。さらに、図形ごとに全被験者の錯視量の平均値と標準偏差を求めた。これらの値を表2と表3、および図2と図3に示す。表2 矢羽の長さの変化と錯視量(mm)15mm30mm35mm45mm平均-16.1 -24.8-23.4 -24.9 標準偏差5.6 5.88.7 9.0 表2 矢羽の長さの変化と錯視量(mm)15mm30mm35mm45mm平均-16.1 -24.8-23.4 -24.9 標準偏差5.6 5.88.7 9.0 表3 矢羽の夾角の変化と錯視量(mm)15度 30度 60度 平均-26.0 -24.8 -17.8 標準偏差7.5 5.8 7.0 表3 矢羽の夾角の変化と錯視量(mm)15度 30度 60度 平均-26.0 -24.8 -17.8 標準偏差7.5 5.8 7.0　表2および図2によると、矢羽の長さが最も短い15mmの場合に最も錯視量が小さく、最も長い45mmの場合に最も錯視量が大きかった。しかし、30mm～45mmの間では、矢羽が長くなれば錯視量が増えるという単純な関係ではなく、錯視量にあまり大きな変化はなかった。　矢羽の長さの4条件について一元配置分散分析を行ったところ、0.1%水準で有意であり(表5)、Tukeyの法による多重比較では15mmの条件と他の3条件の間に5%水準の有意差が認められた。表5 矢羽の長さと錯視量の関係の分散分析平方和df平均平方F矢羽の長さ1169.9333389.9787.034***誤差377 0.21268 55.444全体4940.14571** *: p<.001　表3および図3によれば、15度、30度、60度と、夾角が大きくな	ストループ効果発達臨床学科2年9組Dグループ学籍番号: BxBxxx氏名:白梅香目的本実験では、Stroop(1935)に基づくストループ・カラーワード・テストを用い、文字読みと色名呼称における認知的葛藤状態が反応時間に与える効果を測定し、認知的葛藤の強さと反応時間の関係を検討することを目的とした。ストループ・カラーワード・テストは、色を表す文字がその文字の示す色とは違った色で塗られている。例えば、「赤」という字が青、緑、茶、黒のどれかで書かれている。これらの色文字を、色とは関係なく字を読む場合(文字読み条件)と、字とは関係なく色の名前をいう場合(色名呼称条件)に二通りに用いる。字を読もうとすると色が干渉し、色の名前を言おうすると字が干渉する。そのために、どちらの場合も、二つの異なる反応傾向が妨害しあう状態、即ち、認知的葛藤状態が作り出されるとされている。方法 刺激および用具：ストループ・カラーワード・テスト、ストップウォッチ3種類のカード(黒文字カード、色刺激カード、カラー文字カード)を用いた。どのカードにも、縦横10×10の100 個の漢字または色刺激が並んでいた。黒文字カード：文字読み(黒文字)条件に用いた。色刺激カード：　色名呼称(色刺激)条件に用いた。 カラー文字カード：文字読み(カラー文字)条件と、色名呼称(カラー文字)条件に用いた。被験者：大学生18名。手続き：被験者一人について、次の4条件での試行を行い、所要時間を測定した。文字読み(黒文字)条件： 黒インクで書かれた漢字を読む色名呼称(色刺激)条件： 色を見てその色の名前をいう文字読み(カラー文字)条件： カラー文字の漢字を読む色名呼称(カラー文字)条件： カラー文字の色の名前をいう実験者は次の教示を被験者に与えてから実験を行った。『今から、色の名前を黒い文字で印刷したもの、色そのものを印刷したもの、色の名前を書いた文字が別の色で印刷されたものを使い、文字または色の名前を言ってもらいます。全部で100個の文字または色がありますから、ひとつひとつ指で指しながら、左上から横に読み進めてください。途中を読み飛ばしたり、行が変わるときに行を飛ばしたり、同じ行を2度読んだりしないよう着実に読んでください。できるだけ速く読み進めてもらいますが、急ぐあまり間違えないように気をつけてください。間違いに気づいたら言い直してから読み進めてください。』被験者は、一つ一つ指で指しながら、第1行の左端から最終行の右下まで文字読みまたは色名呼称を行った。順序効果を相殺するため、被験者の半数は「文字読み(黒文字)」、「色名呼称(色刺激)」、「文字読み(カラー文字)」、「色名呼称(カラー文字)」の順番で行い、残りの半数は「色名呼称(色刺激)」、「文字読み(黒文字)」、「色名呼称(カラー文字)」、「文字読み(カラー文字)」の順番で行った。なお、どの被験者についても、最初の2条件つまり「文字読み(黒文字)」と「色名呼称(色刺激)」に関しては、それぞれの測定を始める前に練習試行を最初の1行だけ、文字読みないし色名呼称させ、その後実際の測定試行を行っ

図１ ML文書とST文書の差分表示

実際には100%の類似でありながら、類似度スコアはそれより低くなるという結果について、類似の割合の他に類似部分の絶対量も関係しているかと考えられるので、同一内容レポートの文量を変化させてみた。具体的には、元のレポート内容を繰り返してコピーすることで文量が２倍、及び５倍のレポートを作った。また、逆に元のレポートの一部を削除することで文量が４割（35%〜45%）程度になるレポートを作った。その結果、２倍レポートではMLが89.7、STが89.1、５倍レポートではMLが91.6、STが90.7、４割レポートではMLが83.7、STが80.7となり、類似部分の絶対量が多いほど類似スコアも大きくなる傾向があることがわかった。一般に、この種の類似度評価においては、LCS(Longest Common Subsequence: 最長共通部分列）を求めるアルゴリズムや、レーベンシュタイン距離（Levenshtein distance）(Levenshtein, 1965)、ジャロ・ウィンクラー距離（Jaro-Winkler distance）(Jaro, 1989; Winkler, 1990)を求める計算が使われていると思われるが、いずれも完全一致する文字列の評価は極端なものになる。したがって、WebClassの場合には、そうした方法以外に文の全体量が反映されるような評価方法が用いられていると推測できる。ただ、１万字ないし１万６千字の文書で、完全に内容が同一であってもスコアが100にならない点については違和感が残る。

以下、元の文書に何らかの変更を加えた文書を作成し、元の文書との間で類似度チェックを行ってみた。

③段落の順序変更

元の文書を段落に区切り、その順序を入れ替えた文書を作成した。機械的に行ったので、当然模擬レポートは意味不明なものとなる。ただ、総体としては元の文書と同じ内容となっている。現実の剽窃ではこうした機械的な変更は行わないであろうが、「第一に…」「第二に…」といった論点の順番を入れ替えたりすることは考えられる。

類似度チェックの結果は、ML文書で84.2、ST文書で81.2であった。完全に同じ場合に比べて多少下がる程度で、高いスコアを維持したと言える。

④段落の一部削除

元の文書を段落ごとに一部削除した文書を作成し、元の文書との間で類似度チェックを行った。削除しただけで別の内容を加えてはいないので、削除後の文書は元の文書の真部分集合になる。削除後の文字数はML文書で938文字（元の文書に対して47.5%)、ST文書で1794文字（同じく55.5%）であった。この変更は一部削除という形で行ったものだが、実際には元の文書に対してオリジナルな記述を一部付け加えた場合という逆の形での剽窃を想定している。

類似度チェックの結果は、ML文書で65.6、ST文書で65.8となった。②や③など完全に同じ内容に比べて評価段階が「引用など近似した段落が含まれています」とさらに一段階下がり、グレーゾーンになっている。

⑤段落間に他文書の段落を挿入

これも④と同じく、部分的に剽窃を行い、オリジナルな記述を交えた場合を想定したものである。ML文書、ST文書のそれぞれに別分野レポートの記述を段落単位で挿入し、ML文書に関しては2920文字（47.8%増)、ST文書に関しては4703文字（45.6%増）の文書を作成し、それぞれ元の文書との類似度をチェックした。

結果はML文書が66.3、ST文書が60.7であった。文書による違いが見られるが、④と同程度の判定となった。

⑥段落ごとの一部削除と他文書の挿入

さらに剽窃部分の割合を少なくするために、④と⑤の変更を同時に行った文書を作成した。変更後の文字数はML文書が2029文字（元の2.7%増)、ST文書が2885文字（元の10.7%減）であった。

類似度チェックの結果は、ML文書が51.9、ST文書が51.6となり、同一部分が減ったことに伴いスコアも低くなった。

以上は段落ごとの変更であるが、以下では文字や言葉単位での変更を加えた結果を報告する。

⑦句読点の削除

これは現実の剽窃手法としては考えにくいが、文字単位での変更を機械的に行い、その影響を見る手段として、句読点を単純に削除した文書を作成し、元の文書と類似度をチェックするということを試みた。削除後の文字数はML文書で1859文字（元の94.1%）、ST文書で3073文字（元の95.1%）であり、句読点の占める割合はおよそ5%であった。

類似度スコアはML文書で76.0、ST文書で69.2となった。一応剽窃が疑われる評価段階にはあるものの、単純な変更であり、実質的には同内容である割にはスコアの下がり方が大きいと言えるのではないだろうか。

⑧文末の変更

単純な変更で読み手に与える印象を変える手段として、文末表現を本来の論文形式の「である」調から「です、ます」調に変更してみた。

類似度スコアはML文書が82.7、ST文書が78.7となり、句読点削除ほどスコアは下がらず、②の同一内容のスコアに近い値を維持した。

⑨キーワードの変更

文中のキーワード、すなわち主要な語句を他の類似の語句に変更することで、読み手への印象を変えようとすることは考えられる。ただ、類似の語句への変更を機械的に行うことは難しいので、今回は2文字以上連続する漢字の文字列の順序を逆順にすることで「類似の語句」として、それに置き換えてみた。例えば「標準刺激」は「激刺準標」、「認知的葛藤状態」は「態状藤葛的知認」となる。もちろんこれらはありえない語句になるが、擬似的な類似語と一応考えられるだろう。

類似度スコアはML文書で55.8、ST文書で55.9となり、スコアを減らす効果はかなりあったと言える。同じ文字が使われていても順序を入れ替えただけでかなりの効果があるので、文字の異なる言葉に置き換えられた場合にはさらにスコアが下がるということが考えられる。

⑩文末とキーワードの変更

最後に、⑧の文末表現の変更と⑨のキーワードの変更を同時に行ってみた。結果はML文書で55.8、ST文書で53.9であり、キーワード変更と変わらないか、やや低い程度のスコアで、文末表現の影響はここでもあまりないようだった。

以上の類似度スコアを表2にまとめて示す。

まとめ

以上の検証結果をまとめると、類似度スコアは同一とみなせる表現の総量に主に依存する数値のようである。そして、その数値は、完全に同じ文書でさえ「強い剽窃の疑い」のレベルにようやく達する程度であったり、キーワードを変更するだけで「近似した段落が含まれます」のレベルに落ちてしまったりと、全般に「控えめな」判定となっている。今回、検証に使用したレポートは心理学の実験レポートであり、決まった表現が使われることの多い文書であるが、こうした文書においても控えめな類似度スコアが出るということは、もっと表現に自由度の高いレポートの場合は、さらに低い値となることが予想される。

今回の検証では、段落ごとに大幅な変更をしたり、キーワードを変えたりしても、50程度のスコアは出ていたので、現実的にはスコアが40程度以上あるレポートのペアが見つかった場合には、それらをよく比べてみる必要があるだろう。

表2 比較対象ごとの類似度スコア

比較対象のレポート	ML文書 （ミューラー・リヤーの錯視レポート） 1976文字	ST文書 （ストループ効果レポート） 3231文字
別分野のレポート	11.6	13.6
実験レポート同士（ML対ST）	24.0	
完全同一内容	85.3	82.0
完全同一内容（内容を２倍に）	89.7	89.1
完全同一内容（内容を５倍に）	91.6	90.7
完全同一内容（内容を４割程度に）	83.7	80.7
段落の入れ替え	84.2	81.2
段落ごとに一部削除	65.6 （938文字）	65.8 （1794文字）
段落間に他文書の段落を挿入	66.3 （2920文字）	60.7 （4703文字）
段落ごとの一部削除と他文書の挿入	51.9 （2029文字）	51.6 （2885文字）
句読点削除	76.0 （1859文字）	69.2 （3073文字）
文末を「ですます」調に変更	82.7	78.7
キーワード変更	55.8	55.9
語尾変更とキーワード変更	55.8	53.9

もとより、類似度スコアだけで剽窃の判断はできないわけで、最終的な判断は評価者の吟味に委ねられることになるが、その前段階のチェックとして、WebClassの類似レポート検出機能は実用性があると考えられる。

引用文献

Jaro, M. A. 1989 Advances in record linkage methodology as applied to the 1985 census of Tampa Florida. Journal of the American Statistical Association. 84（406）, 414–420.

倉澤寿之　2016　剽窃検出ソフトWCopyfindの日本語文書への適用について 白梅学園大学・短期大学情報教育研究 第19号, 19-25.

Levenshtein, V. I. 1965 Binary codes capable of correcting deletions, insertions, and reversals. Doklady Akademii Nauk SSSR 163（4）, 845–848.

Winkler, W. E. 1990 String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. Proceedings of the Section on Survey Research Methods. American Statistical Association, 354–359.