K-060

検索エンジンを用いた剽窃レポート発見のための 1 文単位の検索クエリ作成手法

Query selection method of Web search based on each sentence in detecting plagiarism report.

鈴木 啓司† 黒岩 丈介† 小倉 久和† 小高 知宏† 白井 治彦‡ 高橋 勇‡‡

Keiji Suzuki Jousuke Kuroiwa Hisakazu Ogura Tomohiro Odaka Haruhiko Shirai Isamu Takahashi

1 はじめに

我々は，Web サイトからの剽窃レポート発見支援システムと複数の Web ページから剽窃したレポートの発見支援システムの研究を提案してきた [1, 2]．このシステムは，学生が提出したレポートの剽窃元である可能性の高い Web ページを提示し，教師のレポート剽窃チェックを支援するものである．

システムでは，検索エンジンを利用して，レポートから表層的に類似した Web ページを収集する．その際重要となるのが，検索エンジンにリクエストする検索クエリを作成する方法である．文献 [1, 2] では，文章全体から文字列の長い単語 3 つを利用して検索クエリを作成していた．この方法では，その 3 つの中に剽窃元に含まれていない単語があると検索結果に剽窃元が含まれず，Web ページの収集に失敗してしまうことがある．

そこで本研究では．単語の組み合わせを複数作り，検索することで剽窃元を収集できると考えた．また，レポートの剽窃は，複数文を複数箇所から剽窃することが多く見られる．そこで本研究では，特徴から 1 文単位で検索クエリを作成する手法を提案し，実際に学生が提出したレポートで実験を行い，有効性を示すことを研究の目的とする．

2 検索クエリ作成における問題点

検索エンジンは，リクエストに含まれる単語または条件式から検索結果を返す．まず，剽窃レポートから剽窃元の Web ページを発見するには，剽窃元に含まれている単語で検索しなければ剽窃元を含むんだ検索結果を得ることができない．次に，複数の Web ページから剽窃し作成されたレポートの場合，ある剽窃元に含まれている単語のみで検索しなければ適切に検索結果を得ることができない．

後述する検索エンジンの制限から，上位の検索結果に絞り込まなければ検索結果を得ることができない．剽窃レポートは，どの部分が剽窃でどの部分が剽窃でないか判別するのが難しく，複数の Web ページから剽窃されている場合，どの部分が別々の剽窃元から剽窃された部分か判別するののが難しく，その上で検索クエリを作成する必要がある．

検索エンジンについては，Yahoo! デベロッパーネットワークを用いた [3]．このプログラムから検索クエリとクエリの単語選択条件を与えることで，検索結果を XML 形式で取得することが可能である．検索結果は，上位 1000 件まで，1 度のリクエストで 50 件 まで，24 時間で 50000 リクエストまでという制限があること，検索結果にもとづいて Web ページをダウンロードするしなければならないことを考慮する必要がある．

文献 [1, 2] では，文章中の文字列の長い 3 つの単語を利用して検索クエリを作成する．この方法では，剽窃元に含まれない単語があると適切に検索することができず，剽窃元の収集に失敗してしまうという問題がある．

3 1 文単位での検索クエリ作成手法

学習者がレポート作成の際に剽窃する場合，Web ページから文をコピーする特徴がある．このことから，Web サイトから「1 文以上をコピー＆ペースト」したレポートを剽窃行為とする．

そこで，「1 文単位で検索クエリを作成する」方法を以下に示す．

1. 学生のレポート文章を句点で文に分割し，$s_1, s_2, \ldots, s_m$ とする．
2. s_i からそれぞれカタカナ・漢字・アルファベット続きの部分を抜き出す．
3. s_i から抜き出した文字列で長い方から $w_1, w_2, \ldots, w_n$ とする．
4. $w_1 \cdot w_2 \cdot w_3$ を検索クエリとする．
5. (2),(3),(4) を i〜m まで繰り返す．

これにより，文の中の文字列の長いものから 3 つを AND した検索クエリを文の数だけ得ることができる．

4 評価実験

4.1 実験目的

実際に授業で提出された学生レポートにおいて文献 [1], 文献 [2] で利用されている検索クエリの作成方法と本手法とで，剽窃元の検出精度がどう変化するか実験を行った．

4.2 実験条件

実験に使用したレポートは，「計算機システム」の授業で出題されたレポートで，再提出を含め 74 件である．課題のテーマは「DVD は進化途上の光ディスク装置であ

† 福井大学大学院工学研究科
‡ 福井大学工学部
‡‡ 北里大学一般教育部

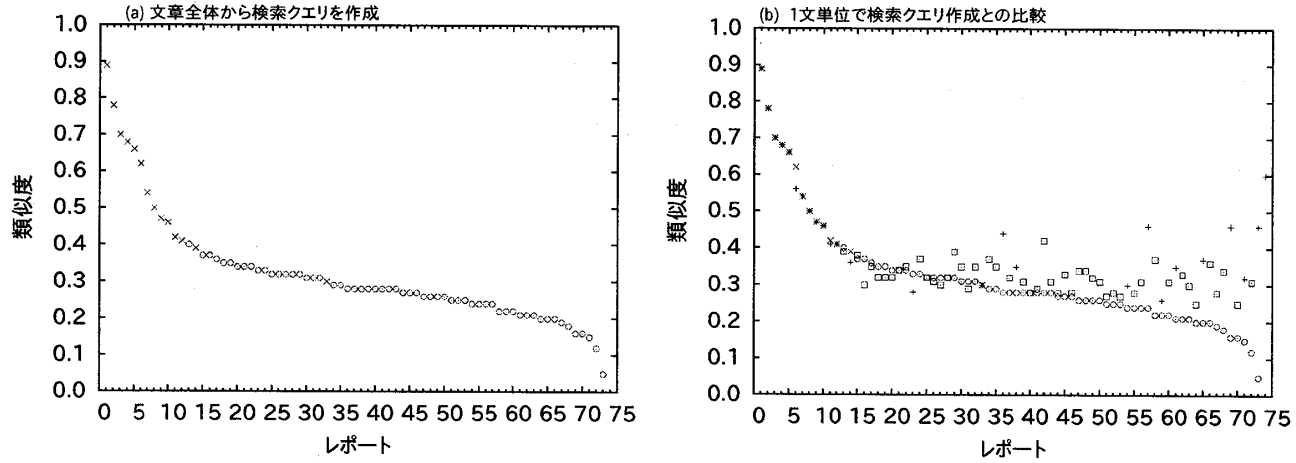

図 1: システムによる剽窃検索の結果

る．最近の動向を調査せよ.」となっている．文字数制限があり 1000 字程度のレポートである．検索エンジンには，Yahoo!デベロッパーネットワークを利用する．

4.3 実験方法

提案手法による，剽窃レポート及びその元となる Web ページの発見手法について，説明する．先ず 3 節で与えた手法によって，対象レポートの各文に対する検索クエリを作成する．その検索クエリを用いて Yahoo! デベロッパーネットワークを用いて Web 検索し，上位 10 件までを剽窃元候補 Web ページとし，候補集合に入れる．同様の手続きをすべての文に対し行い，新たに見つかった Web ページを候補集合に加える．全ての文について Web 検索をし終えたら，n-gram 分布類似度 (n=3) を用いて (文献 [1] 参照)，対象レポートの文章全体と候補集合の剽窃元候補 Web ページの文章間の類似度を算出する．算出された類似度の中で最も高い類似度となる Web ページを剽窃元 Web ページとする．

実験では，実際の講義で学生が課題として提出したレポート 74 件を実験対象とした．このレポートを，文献 [1] で用いている手法（以下従来手法）と提案手法で得られた剽窃元 Web ページについて，目視によって実際に，剽窃かどうか判断し比較を行った．文献 [2] での剽窃元候補 Web ページ発見精度は，1 度目の検索による剽窃元候補 Web ページの収集性能に依存する．つまり，文献 [1] の手法の発見性能である．文献 [1] の手法と比較することで，一文単位で検索クエリを作成することの有効性を評価できると考えた．

4.4 結果

従来手法を用いた結果を図 1(a) に示す．比較のために，従来手法と提案手法の結果を重ねたものを図 1(b) に示す．このグラフは横軸が従来手法で類似度を計算し降順にソートしたレポートの番号で，縦軸は類似度評価値である．また，記号「×」は従来手法を用い主観で剽窃と判断したレポート，「○」は剽窃ではないと判断したレポート，「+」は提案手法を用い剽窃と判断したレポート，「□」は提案手法を用い剽窃ではないと判断したレポートを示す．この実験の結果，従来手法で見付かったレポートは 14 件，提案手法で見付かったレポートが 25 件あった．また，従来手法で見つかっていなかったレポートについて提案手法を用いたことで剽窃元が見つかった場合が 11 件/ 74 件あった．

5 まとめ

本研究では，検索エンジンを用いた剽窃レポートの剽窃元発見に利用する検索クエリの作成方法について提案した．また，検索クエリが剽窃元発見に大きく影響し，今後様々な検索クエリ作成方法と比較を行う必要があると考えられる．また，本研究で用いた類似度評価は n-gram の分布を類似度計算に利用しているため，類似度が低いとき，剽窃と非剽窃の判別が難しい．これは，文ごとに類似度計算を行うことや，文字列の一致を考慮する剽窃評価の方法が必要だと考えられる．

参考文献

[1] 高橋勇，宮川勝年，小高知宏，白井治彦，黒岩丈介，小倉久和．Web サイトからの剽窃レポート発見支援システム．電子情報通信学会論文誌 D, Vol. 90, No. 11, pp. 2989–2999, 2007.

[2] 上野修司，高橋勇，黒岩丈介，白井治彦，小高知宏，小倉久和．複数の Web ページから剽窃したレポートの発見支援システムの実装．情報処理学会研究報告．コンピュータと教育研究会報告, Vol. 2006, No. 130, pp. 41–46, 20061209.

[3] Yahoo!JAPAN. Yahoo!デベロッパーネットワーク. http://developer.yahoo.co.jp/.