コピーレポート検知システムの開発

06H051 竹中康朝

1 はじめに

今日、インターネットの普及により情報の取得は容易になった。しかし容易に情報を仕入れることが可能になり、Web 上からコピーした文章をそのまま載せたコピーレポートを提出する学生が増えてしまった。しかし、コピーレポートは人の目で判別することは大変な労力を要する。しかし放置すれば学生の質が低下するばかりか、学生に対して正当な評価を下せなくなってしまう。そこで、コピーレポートかどうかを自動検知するシステムの開発を行った。本研究では、メールの本文に直接記述する形式のレポートを対象とする。プログラムにより自動的に類似度を調べあげて、Web 上からコピーペーストしたレポートかを判断する。本研究では、メールを受信したときに自動的にプログラムが起動するので、教員の負担を大幅に軽減できる。メールを受信すると、それをトリガーにプログラムが起動する。類似したシステムとして、大阪産業大学工学部情報システム工学科で稼働している TOM が存在する。TOM は類似度の比較対象が他の学生が提出したレポートである。しかしながら、アルゴリズムは TOM と似通っているので、今回開発したシステムは TOM の Web 版と言える。実験の結果、判定をシステムが行うので、教員側への負担は小さいと結論づけた。しかし、類似度判定の精度が高くないので、今後の課題としては、類似度判定の式に改良を加えて精度の向上を図る。

2 開発したシステム

メールを受信するとプログラムが起動する。まず、形態素解析ソフト kakasi を用いてメール本体を分解して文章を単語ごとに分割する。分割された単語を使って、検索エンジンを使って検索する。検索エンジンには Yahoo を利用した。 検索エンジンの選択肢としては、他に GoogleAPI が存在するが、クエリ制限が存在するために Yahoo を利用した。Yahoo の検索結果の HTML から比較対象となる Web ページの URI を取り出した後、取り出した URI を使って Web ページの HTML を取得する。最後に、取得した HTML とメールとを比較して類似度を求め、結果を記述したメールを教員側に送信する。

3 まとめ

本システムを用いてテストを行った。Wikipedia からコピーした文章を貼り付けたメールを送信したところ、75% 類似しているとの結果が出た。100% コピーにも関わらず 75% 類似となっているので、計算式に改良を加える必要がある。既存の TOM との連携を図ることで他の学生が提出したレポートと Web 両方から類似度比較が可能となり、コピーレポートの検知率が向上する。教員への負担は、受信されたメールの内容が Web からのコピーでないかの判断をプログラムが行うので、教員への負担はかなり低いと言える。

3.1 今後の課題

- TOM との連携: 現在稼働している TOM との連携を図ることで、他の学生が書いたレポートと Web 両方から類似度比較を行う。

- URI 抽出機能強化: 検索結果の HTML から URI を取り出すとき、HTML に記述されている URI が一部抜けてしまう場合があるので、全ての URI が表示されるように改良を加える。Yahoo の検索結果の HTML では URI の途中でタブが入っている場合がある。そのタブの存在をプログラムに判断させて除去することで URI の漏れを防ぐ。

- 精度の向上: 類似度の判定の精度を向上させて、コピーレポートを確実に見分けさせる。現在使用している判定式にも改良を加えることで精度が向上させる。