Webページを対象とした不正引用の検出に関する研究

樫山武浩† 田中成典‡ 木下智弘† 増満光† 原川秀哲‡
関西大学大学院総合情報学研究科† 関西大学総合情報学部‡

1．はじめに

近年，Blog や SNS（Social Networking Service）の普及により，ユーザが Web で情報を発信することが容易になった．それに伴い，歌詞や書籍を不正引用した Web ページが増加し，著作者が得るべきだった利益が減少している．そのため，不正引用した Web ページを発見する必要性が高まっている．既存研究では，LCS（Longest Common Subsequence）[1]や N-gram[2]を用いた手法で文章間の類似度を算出し，不正引用した文章を抽出している．LCS を用いた手法[3]では，不正引用した文が他の文章に埋もれていた場合でも類似度を算出することができ，N-gram を用いた手法[4]では，文章の一部に改変がなされた文書間においても類似度を評価できる．しかし，文章間の類似度のみでは，著作権法を考慮していないため，不正引用であるかの判定は不可能である．そこで，本研究では，著作権法の引用形式[5]を考慮し，著作物を不正引用した Web ページの検出を目的とする．

2．システムの概要

本研究では，著作物を不正引用した Web ページの判定手法を提案する．本システムの概要を図1に示す．本システムは，1）Web ページ収集機能，2）類似 Web ページ抽出機能，3）不正引用判定機能の3つの機能で構成される．入力データは，新聞記事や歌詞など文章で作られた著作物とし，出力データは不正引用した Web ページとする．

2．1　Webページ収集機能

本機能では，入力した著作物に類似する可能性がある Web ページ群を収集する．まず，Web ページの収集手法として，著作物の文章の先頭から文節を取得する．次に，取得した順番に2つの文節を取り出し，1つの検索語として Web ページの収集を行う．そして，2つの文節が取得できなくなるまで Web ページの収集を行う．

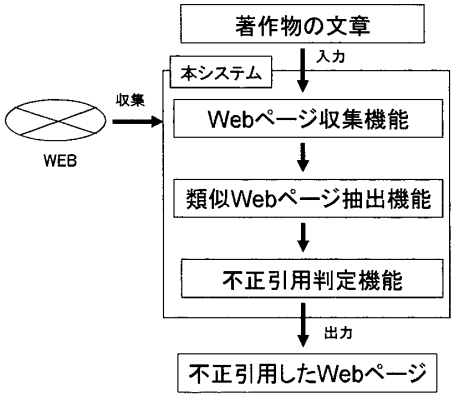

図1　本システムの概要

2．2　類似 Web ページ抽出機能

本機能では，著作物を引用した可能性が高い Web ページだけを抽出する．Web ページの抽出は，著作物の文を改変することや冗長させることによる引用を考慮し，著作物の文章と Web ページの文章を文単位に分割する．そして，文同士を形態素単位における N-gram の手法で比較を行い，著作物と類似性の高い Web ページだけの抽出を行う．

2．3　不正引用判定機能

本機能では，類似 Web ページ抽出機能で抽出した Web ページに対し，不正引用であるかの判定を行う．本研究で考慮する著作権法の引用形式は，“引用部分を明確にすること”と“引用元の著作者と書物名を明確にすること”とする．まず，Web ページの引用箇所を抽出するため，著作物の文章と Web ページの文章において，N-gram の手法を用いて文字単位で類似する文の抽出を行う．次に，抽出した文が，Web ページ内で集中して出現する箇所を引用箇所として抽出する．最後に，引用形式を確認するため，HTML タグの有無と引用元の著作者と書物名の記述の有無を確認し，確認が取れない場合，Web ページを不正引用であると判定する．

Research for Detecting Illegal Quoted Web Articles
† Takehiro Kashiyama, Tomohiro Kinoshita, Hikaru Masumitsu
Graduate School of Informatics, Kansai University, 2-1-1 Ryouzenji-cho Takatsuki-shi, Osaka 569-1095, Japan
‡ Shigenori Tanaka, Hideaki Harakawa
Faculty of Informatics, Kansai University, 2-1-1 Ryouzenji-cho Takatsuki-shi, Osaka 569-1095, Japan

3．システムの実証実験と考察

実証実験では，本研究で提案した手法が，不正引用した Web ページを検出できるか確認し，本システムの有用性を実証する．

3．1　実証実験

実証実験では，著作物として新聞記事６件と歌詞６件を本システムに入力し，著作物の種類ごとにおける本システムの有用性を実証する．入力する新聞記事は，Yahoo! ニュースのアクセスランキングから６件のニュース記事を無作為に選出した．また，入力する歌詞は，うたまっぷ.com と Uta-Net の人気歌詞ランキングから無作為に３件ずつを選出した．本システムに著作物を入力し，本手法において不正引用として判定された Web ページを抽出した．そして，抽出された Web ページを目視で確認し，正確性と網羅性を示す指標として適合率と再現率を算出した．そして，総合的な有用性を評価する指標として，適合率と再現率の調和平均値である F 値を算出した．

3．2　結果と考察

本システムに入力した新聞記事の例を図２に示し，本システムが不正引用であると誤判定した Web ページの文章の例を図３に示す．新聞記事において誤判定される原因として，新聞社各社が同じ話題の記事を配信する場合があり，他社の新聞記事を不正引用と誤判定されるためである．

本研究の実験結果を表１に示す．実験結果より，どの著作物においても再現率が高くなっていることが確認できる．これは，不正引用した Web ページに対する誤判定が少ないと言える．しかし，適合率の結果では再現率と比較して小さい値となった．これは，正しく引用した Web ページや類似性が低い Web ページが不正引用として判定されるためである．原因としては，文字や記号などの HTML タグ以外の表現で“引用部分を明確にすること”の形式を構成している場合と引用箇所の誤抽出や抽出漏れにより不正引用であるかの判定が困難であると考えられる．

4．おわりに

本研究では，著作権法の引用形式を考慮し，不正引用した Web ページの判定を行った．本手法では，Web ページの引用箇所を抽出し，HTML タグと引用元の著作者と書物名の有無を基にし，Web ページを不正引用であるかどうかを判定した．実証実験では，不正引用した Web ページに対する誤判定が少ないことから本手法の有用性を示した．今後は，適合率を向上させるため，Web ページにおける引用箇所の抽出精度の向上と HTML タグ以外の文字や記号の考慮をした手法を調査する予定である．また，著作権法の引用形式の１つである“引用者の書く内容が量的にも質的にも主であり，引用部分が従であること”の形式を考慮した手法も提案する予定である．

著作者：読売新聞社
著作物名：イスラエル軍、ガザ攻撃で「非人道兵器」使用か
　イスラエル軍によるパレスチナ自治区ガザ攻撃で、ＡＰ通信は１１日、現地住民の証言として、同国軍がイスラエル境界に近いフーザ村で住宅に向けて、「非人道兵器」として

図２ 本システムに入力した新聞記事

イスラエル軍がパレスチナ自治区ガザへの攻撃で、非人道兵器とされる「白リン弾」を使用しているとの疑惑が浮上している。白リン弾によるとみられる民間負傷者が報告されている。

著作物名：深まる白リン弾疑惑
時事通信社の記事から引用

図３ 本システムが誤判定した例

表１　実験結果

	再現率	適合率	F 値
新聞記事	0.84	0.62	0.71
歌詞	0.73	0.67	0.69

参考文献

[1] James, W. Hunt, Thomas, G. Szymanski : A Fast Algorithm for Computing Longest Common Subsequences, Communications of the ACM, Vol.20, No.5, pp.350-353, 1977.5.

[2] Matsuura, T, Kanada, Y. : Extraction of Authors' Characteristics from Japanese Modern Sentences via N-gram Distribution, Lecture Notes in Computer Science, Springer Berlin / Heidelberg, Vol.1967, pp.315-319, 2000.1.

[3] 田代崇，上田高徳，堀泰祐，平手勇宇，山名早人：Web ページを対象とした著作権違反自動検知システム，データベース・システム研究会報告，情報処理学会，Vol.2006，No.78，pp.27-33，2006.07.

[4] 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システム，電子情報通信学会論文誌，電子情報通信学会，Vol.J90-D，No.11，pp.2989-2999，2007.11.

[5] 文化庁長官官房著作権課：著作権テキスト，文化庁，2008.