著作権侵害文書検出のための要約文発見手法

A Discovering Method of Summarized Documents for Detecting Copyright Infringement

雲居 玄道[†] 濱田 弥[††] 坂口 朋章[††] 八木 秀樹[†††] 平澤 茂一[†]
Gendo Kumoi[†] Hisashi Hamada[††] Tomoaki Sakaguchi[††] Hideki Yagi[†††] Shigeichi Hirasawa[†]

[†] 早稲田大学創造理工学部
[††] 早稲田大学大学院創造理工学研究科
[†††] 早稲田大学メディアネットワークセンター

[†] School of Creative Science and Engineering,Waseda University
[††] Graduate School of Creative Science and Engineering,Waseda University
[†††] Media Network Center, Waseda University

要旨

近年，インターネットの普及によって，デジタル文書の著作権侵害が深刻な問題となっている．この問題の1つに，著作物を著作権所有者の許可なく要約し，これを公開することがあげられる．本研究では，著作権の保護対象となる文書から，要約されたデジタル文書の発見する手法を提案する．この手法は，要約文は原著作物に依拠して作成されているという観点から2文書間の最大連続文節数によって判定する手法と，発見対象文書中の全ての単語に対し著作物文書に含まれている単語数の割合によって判定する手法である．新聞記事とそれを元に人手で要約されたデジタル文書を用いて評価実験を行い，提案システムの有効性を示す．

Abstract

Copyright infringement of digital documents has become a serious problem recently due to wide use of internet. One of this problems is unauthorized actions of creating summaries of documents without author's permission. In this paper, it is suggested how to detect unauthorized summaries of the digital documents requiring protection against copyright infringement. The concrete procedure of applying this detecting system is demonstrated. We suggest this method considering the fact that summaries are created based on the original writings. One of our suggested techniques employs the number of the longest serial phrases shared between two documents, and the other adopts the ratio of the words included in the original writing to the number of all the words of the object document. In order to prove effectiveness of the proposed methods, we show the experimental results by using newspaper articles and the summaries of them manually created.

1 はじめに

近年，Eメールやインターネットの普及と共にデジタル文書が急増している．メールマガジンやblogなどを通じて多くのユーザが安易に情報を発信するようになり，それに伴い著作物の著作権侵害が深刻な問題となっている．著作権侵害のひとつに著作物の要約があげられる．

現在，新聞や雑誌，書籍に至るまで多くの文書が存在しており，その全てを読むことは不可能である．しかし，内容を把握しておく必要性のある文書も多くあり，それらを要約し配信するWebページやメールマガジンが数多く存在している．また，最近では，アフェリエイト[1]の流行により，書籍の要約をblog等で紹介する行為も増加している．従って，著作物の要約を自動的に検出することは，著作権侵害文書検出において重要な位置を占めつつある．

従来，著作権侵害のひとつである剽窃を発見する手法として，学生レポートに関する研究[1]やR. W. Irvingによって提案されたアルゴリズム[2]などがある．しかし，人手要約[2]では様々な換言や複文の単文化などが施されており，従来の剽窃を発見する手法では対応が困難である．また，人手要約では，原著作物の形式によって様々な要約手法が用いられ，また語の選択もバラツキが大きいため，自動要約の手法によって対応させることも困難である．

本研究では，人手要約に関するモデル化を行い，そのモデルに基づいて，原著作物から要約された文書を発見する手法を提案する．この手法を用いて，デジタル文書集合からの要約された文書の発見を支援する．新聞記事に対する要約文書に適用し，有効性を示す．

[1] Webサイトやメールマガジンなどが企業サイトへリンクを張り，閲覧者がそのリンクを経由して商品を購入したりすると，リンク元サイトに報酬が支払われるという広告手法．

[2] 本研究では，人手による要約を人手要約と呼ぶ．

2 要約の著作権侵害

2.1 著作物とは

著作権法2条1項2号には，著作物性とは「思想又は感情を創作的に表現したものであって，文芸，学術，美術又は音楽の範囲に属するものをいう」とあり，一般的な書籍については著作物性が存在することは疑いようがない．一方，新聞の著作物性については，「新聞記事は，いずれも事実の伝達にすぎない雑報や時事の報道に止まるものではなく，その盛り込む事項の選択，報道事実や論理の展開の仕方，文章表現等に創作性が認められる著作物である」という判例[3]があり，著作物と認められている．従って，書籍・新聞ともに著作権侵害から保護する必要がある．

2.2 要約と翻案権

著作者の権利として，著作権法27条には，「著作者は，その著作物を翻訳し，編曲し，若しくは変形し，又は脚色し，映画化し，その他翻案する権利を専有する」とある．また，翻案にあたる要約として，「原著作物を読まなくても原著作物に表現された思想，感情の主要な部分を認識させる内容を有しているもの」という判例[3]がある．新聞記事の一般的な要約はこれに該当し，要約は翻案権を侵害するため，著作権侵害となる．書籍に対する判例[4]も存在し，書籍の場合も，たとえ一部分の要約であっても，その内容を理解させるものであれば，著作権侵害にあたると考えられる．

近年，blogなどインターネット上に著作権者の許可無く，文書の要約を掲載する事例が多数見受けられる．本研究では，著作権により保護される元の文書（著作物文書）を入力とし，その要約文書（発見対象文書）を自動的にデータベースから発見する手法を提案する．

2.3 要約発見手法の要件

本研究の目的は，要約されたデジタル文書の発見を支援する手法である．そのため，要約文書候補を提示することが目的である．そして，その候補にデータベース中に含まれる要約文書が全て含まれていることを第一に捉える．その上で，候補の数が最小になるようにする．

3 提案手法

翻案権の侵害となる要約の第一の前提条件は，「原著作物に依拠して作成」と示されている[3]．本研究ではこの原著作物に依拠している点を考慮に入れ，要約の作成について以下の2点を仮定する．

1. 人手要約において原著作物の文節の出現順序は，一定数，保持される
2. 人手要約において文中の出現単語は，一定率，原著作物に含まれる

[3]東京地判平成6年1月18日知的裁集26巻1号114頁.
[4]東京地判平成13年12月3日判時1768号116頁.

以下にこの仮定に基づいて，要約された文書を発見する2つの手法を示し，これを組み合わせた手法を提案する．

3.1 手法1

3.1.1 概要

手法1では，入力となる著作物文書とデータベース中の発見対象文書を比較し，連続する文節数の最大値を取計算する．その値があらかじめ定めた定数以上であれば，その発見対象文書を著作物文書から作成された要約文書と判定する局所的な検出アルゴリズムである．要約文検出の目的から，よく知られたSmith-Waterman アルゴリズムより厳しい条件を課している．

3.1.2 アルゴリズム

著作物文書を D_i，発見対象文書を d_j とする．データベース中の全ての発見対象文書 d_j に対し，以下の操作を行う．

[局所的要約文検出アルゴリズム]

Step1 構文解析器によって，形態素解析済みの D_i, d_j を文節に分割する．

Step2 文節に分割済みの D_i, d_j に対し，文の要素となる，名詞・形容詞・形容動詞・動詞を単語として抽出する．ここで，文書 D_i の文節数を e_i，k 番目の文節を $C_{i,k}$，文節内の m 番目の単語を $W_m^{(i,k)}$ とし，文書 d_j の文節数を f_j，l 番目の文節を $c_{j,l}$，文節内の n 番目の単語を $w_n^{(j,l)}$ とする．

Step3 $k := 1, l := 1, q := 0, k' := 0$ とする．

Step4 任意の $C_{i,k}$，$c_{j,l}$ に対し，

$$(\forall m, n)\ W_m^{(i,k)} = w_n^{(j,l)} \tag{1}$$

となる単語を探索し，このような単語が1つでも存在した場合を一致とし Step5 へ．存在しない場合を不一致とし Step6 へ．

Step5 $N_{i,j,q} := N_{i,j,q} + 1, l := l + 1, k' = 0$ ならば $k' := k, k := k + 1$ とし，Step4 へ．

Step5 $l < f_j$ ならば，$l := l + 1, k := k', q := q + 1$ とし，Step4 へ．

$l = f_j$ かつ $k < e_j$ ならば，$k := k' + 1, k' = 0, l := 1, q := q + 1$ とし，Step4 へ．

$k = e_i$ ならば，Step7 へ．

Step7

$$K_{i,j} = \max_q N_{i,j,q} \tag{2}$$

とする．ある定数 $\alpha_1 (> 0)$ に対し，$K_{i,j} \geq \alpha_1$ となるものを要約文書と判定する．

■

3.2 手法2

3.2.1 概要

手法2では，著作物文書と発見対象文書を比較し，発見対象文書中の全ての単語に対して著作物文書に含ま

れている単語の割合を計算する．その割合が一定率以上であれば，その発見対象文書を著作物から作成された要約文書と判定する大域的検出アルゴリズムである．

手法1では，連続する文節数の最大値を取るため，発見対象文書中の一部が一致していると要約文書と判定するが，この手法2を用いることにより，要約文書全体が著作物から作成されたものかを判定できる．

3.2.2 アルゴリズム

著作物文書を D_i，発見対象文書を d_j とする．

[大域的要約文書検出アルゴリズム]

Step1 形態素解析済みの D_i, d_j に対し，文の要素となる，名詞・形容詞・形容動詞・動詞を単語として抽出する．

Step2 発見対象文書中の総単語数を n_j，各単語を $y_r, r = 1, 2, \cdots, n_j$，とし，著作物文書中の単語集合を W_i としたとき，

$$P_{i,j} = \frac{\sum_{r=1}^{n_j} count_r}{n_j}. \tag{3}$$

とする．ここで，

$$count_r = \begin{cases} 1, & \text{if} \quad y_r \in W_i \\ 0, & \text{if} \quad \text{otherwise.} \end{cases}$$

この $P_{i,j}$ を発見対象文書中の単語の著作物文書への依存率とする．本研究では，ある定数 $\alpha_2 (> 0)$ に対し，$P_{i,j} \geq \alpha_2$ となるものを要約文書と判定する．

■

3.3 提案手法

要約文書の発見手法として，提案した手法1，手法2を組み合わせた手法を用いる．この手法は，手法2で要約文書と判定されたものを手法1を用いて絞り込む手法である．

4 評価実験

本研究で提案したシステムを実装し，評価実験を行った．本実験では，著作物文書 D_i として新聞記事，発見対象文書 d_j として新聞記事を人手要約したものを用いる．

4.1 実験データと評価手法

4.1.1 実験データ

著作物文書 要約文書の元となった記事として，[9]～[11]から人手で収集した朝日，読売，日経の3紙の記事693件を用いる．

発見対象文書 宗教関連語が1つでも含まれる新聞記事の要約文書をメールマガジン[5]より，2006年7月3日～2007年7月10日の期間に発行された5,023件を用いる．

[5]浄土真宗本願寺派　本願寺宗務首都圏センター　教学伝道研究センター　本願寺教学伝道研究所　東京支所発行

4.1.2 実験方法

ここでは，予備実験より $\alpha_1 = 4, \alpha_2 = 0.5$ とする．文書の形態素解析にはJUMAN[8]を構文解析にはKNP[7]を用いる．また，実験には，5,023件の要約文書から80件をランダムに選び各手法で実験を行い，それを80回繰り返す．

4.1.3 評価方法

評価尺度には，精度，再現率，F値を用いる．それぞれの定義は以下の通りである．

$$\text{精度} \quad p = \frac{R}{N}, \tag{4}$$

$$\text{再現率} \quad q = \frac{R}{C}, \tag{5}$$

$$F\text{値} \quad F = \frac{2 \cdot p \cdot r}{p + r}. \tag{6}$$

$$\begin{pmatrix} R: & \text{発見された正解要約文書数} \\ N: & \text{発見された総要約文書数} \\ C: & \text{総要約文書数} \end{pmatrix}$$

これら R, N, C は，80回の繰り返しの平均値を用いる．

4.2 実験結果と考察

4.2.1 実験結果

ここで各手法に対して実験を行った結果を表1に示す．

4.2.2 考察

(1) 手法1は，予備実験より $\alpha_1 = 4$ と決定し実験した．これは，新聞記事においては類似の言い回しが数多くあり，連続する3文節以下では，著作物から依拠してつくられたとは言い難くなるためである．しかし，発見対象文書中の一部が一致する文書も要約文書と判定するため，精度が0.34と低くなる．

(2) 手法2は，$\alpha_2 = 0.5$ としたため，発見対象文書中の半数の単語が一致した場合要約文書と判定される．対象文書が新聞記事であり，時事問題などでは使われている単語に類似性があるため，精度が0.027と大幅に低くなる．

(3) 手法1と手法2を組み合わせることによって，手法1の再現率を保ちつつ，精度を大幅に上げることができたといえる．これは，発見対象文書を局所的な視点および大域的な視点の両方からみることにより，より正確な要約文書判定ができたためと考えられる．

(4) 提案手法の計算量は，手法1は，$e_i \cdot f_j$ に比例する．手法2は，n_j に比例する．

4.2.3 結果例

以下に，発見誤り（要約文書と判定されたが，実際には要約文書でないもの）と発見見逃し（要約文書と

表 1: 実験結果

	精度	再現率	F 値	R	N	C
手法 1	0.349	0.889	0.501	9.675	27.725	10.888
手法 2	0.027	1.000	0.053	10.913	398.200	10.913
提案手法	0.528	0.869	0.657	9.488	17.975	10.913

して判定しなければならないが，判定されなかったもの）の例を示す．

発見誤り

新聞記事 "靖国参拝 喉かわしい首相の論法"（8/4 朝日朝刊 3 面・社説）
私を批判するマスコミや識者の意見を突き詰めていくと、中国が反対しているから靖国参拝はやめた方がいい、中国の嫌がることはしない方がいいということになる

要約文書 "「毎年靖国参拝、私の思い」小泉首相メルマガ、「15日」へ地ならしか"（8/3 朝日朝刊 2 面）
突き詰めると、中国の嫌がることはしない方がいいということになる ように思えてならない

発見見逃し

"創価学会 秋谷会長が 6 期目"（7/7 朝日朝刊 4 面）

新聞記事 創価学会，秋谷会長が 6 期目
創価学会は 6 日，東京・信濃町の学会本部で会長選出委員会を開き，秋谷栄之助会長（75）の再任を決めた．秋谷氏は 81 年に会長に就任し，今回で 6 期目．任期は 5 年．

要約文書 「創価学会 秋谷会長が 6 期目」と題する記事．創価学会は 6 日，学会本部で会長選出委員会を開き，同会長の再任が決定した．（任期 5 年）今回で 6 期目となる．

社説はそれ以前の記事のまとめて意見を述べるという要素が強いため，それ以前の記事と全く同一の文が出現する場合がある．このため，多くの発見誤りは，上に示した例のように社説に対して生じている．また，発見見逃しのものは，手法 1 を適用した場合にのみ見られるが，それは，単語の換言・削除に加え，語順の入替が行われていたため，発見できなかったものと考えられる．

5 まとめと今後の課題

本論文では，要約文は原著作物に依拠して作成されているという仮定から要約された文書を発見する手法を提案した．提案手法は，2 文書間の最大連続文節数によって判定する手法（手法 1）と，発見対象文書中の全ての単語に対し著作物文書に含まれている単語の割合によって判定する手法（手法 2）の組み合わせである．前者は部分的な一致，後者は要約文書全体の一致を計るという視点に基づいている．

提案手法の有効性を示すため，新聞記事とその要約文書を用いた評価実験を行った．実験結果から，どちらか一方だけでは，要約文書でないが誤って要約文書と判定される場合が多く，精度が落ちてしまうことが分かった．そこで，2 つの手法を組み合わせることによって，部分的な一致，要約文書全体の一致という視点を併せ持った手法となり，再現率を保ったまま精度を高めることができた．以上より，本提案手法は，著作権侵害文書検出のために原著作物から要約文書を発見する手法として有効な手段であるといえる．

今後の課題として，新聞記事の対象とする記事を宗教関連語が含まれないトピックも含めて実験を行うことが挙げられる．また，インターネット上などでは，書籍の要約も問題である．そのため，評価実験の対象を書籍にまで広げて評価する必要がある．それと同時に精度を向上させるため，要約文書と判定する新しい観点が必要となる．

謝辞

著者の一人雲居は日頃評価を頂く早稲田大学創造理工学部経営システム工学科石田崇助手はじめ平澤研究室の博士課程，修士課程の方々に感謝いたします．

本研究の一部は，（財）電気通信普及財団の研究助成による．

参考文献

[1] 太田貴久，増山繁，"学生レポート採点支援の為のレポート類似部分発見手法"，信学技報，NLC2005-112，pp. 37-42, 2006.
[2] R. W. Irving，"Plagiarism and collusion detection using the Smith-Waterman algorithm"，Technical Report164，Dept. of Computing Science，University of Glasgow，2004.
[3] 大竹清敬，岡本大吾，児玉充，増山繁，"重要文抽出，自由作成要約に対応した新聞記事要約システム YELLOW"，情報処理学会研究報告，2001-F1-63-17，pp. 129-136, 2001.
[4] 相良直樹，砂山渡，谷内田真彦，"サブトピックを考慮した重要文抽出による報知的要約作成"，電子情報通信学会論文誌，Vol. J90-D，No. 2，pp. 427-440, 2007.
[5] 菅野 和夫，小早川 光郎，江頭 憲治郎，西田 典之，"ポケット六法〈平成 19 年版〉"，有斐閣，2006.
[6] 斉藤博，半田正夫，"著作権判例百選"，別冊ジュリスト 157 号，有斐閣，2001.
[7] 黒橋禎夫，"日本語構文解析システム KNP"，http://www.kc.t.u-tokyo.ac.jp/
[8] 黒橋禎夫，"日本語形態素解析システム JUMAN"，http://www.kc.t.u-tokyo.ac.jp/
[9] 日経テレコン 21（日経新聞記事データベース）
[10] ヨミダス文書館（読売新聞記事データベース）
[11] 聞蔵 2 ビジュアル（朝日新聞記事データベース）