Vol.2009-NL-194 No.1
2009/11/16

# Webサイトからの盗作の自動検出システム

馬　　青[†1]　松山秀人[†2]　村田真樹[†3]

Webサイトからの盗作の自動検出システムを構築した．盗作検出に必要な文書類似度の計算には，従来の文字ベースn-gram手法に加え，新たに単語ベースn-gram手法，(同義語変換への対処を想定した）シソーラス手法，さらに文字ベースまたは単語ベースn-gramとシソーラスの利用を組み合わせた手法を提案した．これらの手法について，盗作元とされるWebページに，文末変換・順序変更・文削除・類義語変換によるそれぞれの改変とそれらの混合改変を加えることにより作成した疑似盗作データと被験者が作成した課題レポートを用いた類似度の評価実験を行った．その結果，盗作の程度変化の検出や盗作かオリジナルかの区分けには提案手法である単語ベース手法及び組み合わせ手法の方が従来の文字ベース手法より優れていることがわかった．

# Automatic plagiarism detection from Web sites

Qing Ma,[†1] Hideto Matsuyama[†2]
and Masaki Murata[†3]

We have developed a system that can detect plagiarism from Web sites. To measure document similarity, which is necessary for plagiarism detection, we proposed several new methods that demonstrate a higher performance than the traditional character-based n-gram method. These include a word-based n-gram method, a thesaurus method, and a combination n-gram/thesaurus method. To evaluate the effectiveness of the proposed methods, we performed computer experiments using pseudo plagiarism data and reports made by test subjects. The results demonstrate that the performance of the proposed methods was better than that of the conventional one.

## 1. はじめに

様々なほしい情報をインターネットから手軽に入手できる高度情報化社会の今日において，新聞記者の記事盗用事件に関する報道がよく耳にする．また，大学では，学生が課題レポート（または卒論の一部など）としてWebサイトからコピーしたものをそのまま（あるいは微少な変更で）提出してくる不正行為も身近によく見受けられる．受講者数がたとえば100名を超えていれば1課題につき提出されるレポートの数が100を超えることになり，コピー元となる可能性の高いWebサイトは大量に存在していることを考えれば，盗作かどうかを教員の人手でチェックするのはたいへんな労力のかかる仕事であることがわかる．

このような背景下で高橋らはWebサイトからの盗作検出の支援システムの開発を行っている[1)]．盗作を検出するためのコア技術は文書間の類似度計算であり，これまで盗作検出を念頭においた文書間の類似度計算に関する研究は多数なされていた（たとえば，文献2),3),4)）．ただし，これらの研究は学習者間のレポートの類似度を評価することに留まっている．それに対し，前述の高橋らの支援システムでは小高らの手法[2)]をベースにレポートとWebページ間の類似度評価を行うことによりWebから盗作の検出を可能とした．なお，より多くの先行研究の文献は高橋らの論文[1)]から参照でき，高橋らの研究との比較も同論文に詳しく述べられている．

高橋らの研究には，しかしながら，以下のような問題が存在する．盗作については，盗作元の文書に対し，文末変換・順序変更・文削除という改変は想定していたが，ごまかしによく使われるであろう，特定のいくつかの単語を似た意味の単語に置き換える，つまり類義語変換という改変は考慮に入れていなかった．この類の改変を検出するためには，類義語辞書の利用が必要となるが，高橋らの手法では，類似語辞書の利用はできない．

本研究では，独自のWebサイトからの盗作の自動検出システムを構築した．盗作検出に用いる文書類似度計算に，従来の文字ベースのn-gram手法に加え，新たに単語ベースn-gram手法，(同義語変換への対処を想定した）シソーラス手法，さらに文字ベースまたは単語ベースn-gramとシソーラスの利用を組み合わせた手法を提案した．これらの手法につ

†1 龍谷大学
Ryukoku University
†2 大日本スクリーン製造（株）
Dainippon Screen MFG. Co., Ltd.
†3 （独）情報通信研究機構
National Institute of Information and Communications Technology

いて，盗作元とされる Web ページに，文末変換・順序変更・文削除・類義語変換によるそれぞれの改変とそれらの混合改変を加えることにより作成した疑似盗作データと被験者が指示通りに作成した課題レポートを用いた類似度の評価実験を行った．その結果，盗作の程度変化の検出や盗作かオリジナルかの区分けには提案手法である単語ベース手法と組み合わせ手法の方が従来の文字ベース手法より優れていることがわかった．

## 2. 盗作の定義

本研究で取り扱う盗作を以下のように定義する．

(1) Web ページからのコピーそのもの（=コピー元）
(2) 順序変更
コピー元の文書に対し文の順序を入れ替えて得られたもの
(3) 文削除
コピー元の文書に対し文全体または文中の一部分を削除して得られたもの
(4) 同義語変換
コピー元の文書に対し，類義語（たとえば「使用」と「利用」）を相互に変換して得られたもの
(5) 文末変換
コピー元の文書に対し文末表現の「です・ます調」の記述と「だ・である調」の記述を相互に変換して得られたもの
(6) 複合改変
コピー元の文書に対し，上記 (2)〜(5) のすべての操作を行って得られたもの．

## 3. システムの概要

盗作判定の対象文書（以降「対象文書」と略す）が与えられたとき，その文書から重要語を抽出し，それらをキーワードに Google で Web 検索を行う．検索で得られた上位 20 件の Web ページに対し，文字コードの統一処理や html タグの除去処理を行い，コピー元の候補文書とする．次に，対象文書とコピー元の候補文書との間の文書間の類似度を次節に述べる方法で計算する．最後に，もっとも類似度の高い候補文書を類似度とともにユーザに提示する．

Web 検索において適切なキーワードを抽出して用いることが重要である．高橋らの研究ではレポート内から最長単語長の上位 3 位までの単語を抽出して検索のキーワードとして用いた．しかし，このような検索方法は同義語変換を行った盗作に対し，それらのキーワードが同義語に変換されていることが考えられるため，適切な Web ページを発見できない可能性がある．本研究ではキーワードを提出してきたレポートの中からではなく，教員などによって与えられたレポート課題そのものから抽出して用いている．たとえば「ちりめんじゃことしらすの違いについて」というレポート課題であれば，その課題から内容語「ちりめんじゃこ」と「しらす」と「違い」を取り出しキーワードとして Google で AND 検索を行う．

## 4. 文書間の類似度計算

### 4.1 n-gram 手法

#### 4.1.1 文字ベース n-gram 手法

高橋らが提案した手法で，文書と文書をそれぞれ単一の文字列としてとらえ，それらを 1 文字ずつずらしながら分割して得た n-gram パターンの分布がどれだけ似ているかで文書間の類似度を測るものである[⋆1]．

ここで，対象文書を $P$，コピー元の候補文書を $Q$，対象文書に出現する n-gram パターンを $X = (x_1, \cdots, x_n)$ とすれば，文書間の類似度 $R$ は以下で計算する．

$$R = 1 - \frac{1}{K} \sum_{i=1}^{K} \left( \frac{P(X_i) - Q(X_i)}{P(X_i) + Q(X_i)} \right)^2 \quad (1)$$

ただし，

$K$: $P$ における n-gram パターンの種類数（異なり総数）
$X_i$: $P$ の $i$ 番目の n-gram パターン
$P(X_i)$: $P$ における $X_i$ の出現頻度
$Q(X_i)$: $Q$ における $X_i$ の出現頻度

#### 4.1.2 単語ベース n-gram 手法

文書間の類似度の計算には，長い複合語や専門用語などは本来，まとまったものとして扱うのが望ましい．しかし文字ベース n-gram 手法を用いると，それらを他と無差別に細かく切ってしまい，まったく意味のない文字列になってしまう．たとえば，「自然言語処理」

⋆1 ただし，高橋らの研究では n を 3 に固定して評価を行ったが，本研究では n を 3〜5 に変化させながら評価実験を行っている．

は，高橋らの手法では「自然言」「然言語」「言語処」「語処理」という四つの意味をなさない tri-gram パターンに分割される．このような問題点を改善するために，本研究では，提案手法の１つとして単語ベース n-gram 手法を導入した．本手法は，形態素解析ツール茶筌で単語分割し，文字の代わりに単語の n-gram パターンを形成し，それらの分布がどれだけ似ているかで文書間の類似度を測るものである．すなわち，対象文書 $P$ とコピー元の候補文書 $Q$ との類似度 $R$ は式 (1) で計算する．ただし，n-gram パターンは単語ベースのものである．

### 4.2 シソーラス手法

高橋らの先行研究では類義語変換に関する実験を行っていなかったため，文字ベース手法がどれだけ類義語変換に強いかがわからない．本手法は，類義語変換しても盗作として検出できるように導入した手法である．本手法ではまず，文書を形態素解析し，内容語である名詞，動詞，形容詞，形容動詞の原形を取り出す（ただし，大きな意味をなさない場合が多いと思われる「する」などサ変活用動詞と一文字単語を除外する）．次に得られた内容語に対し分類語彙表[5] を用いて分類番号（以降，これを「概念値」と呼ぶ）を付与する．文書間の類似度は（n-gram パターンの代わりに）概念値の分布がどれだけ似ているかで測る．すなわち，対象文書 $P$ とコピー元の候補文書 $Q$ との類似度 $R$ は式 (1) で計算する．ただし，「n-gram パターン」は「概念値」で置き換えたものとする．

### 4.3 組み合わせ手法

本手法は，文の入れ替えなど表層的な変更に強い（類似度が高く出る）n-gram ベース手法と，類義語変換に強いシソーラス手法を組み合わせたものである．この手法による文書間の類似度の計測の基本的な考え方は，対象文書 $P$ の単語をコピー元の候補文書 $Q$ の単語で意味的類似の条件で置き換えた上で文書間の類似度を計算することである．具体的な処理手順は以下のとおりである[⋆1]．

**手順１**　文書 $P$ と $Q$ を形態素解析し，単語分割を行うとともに名詞，動詞（サ変活用動詞を除く），形容詞，形容動詞以外のものを取り除く．

**手順２**　各単語に分類語彙表を用いて概念値を付与する．

**手順３**　文書 $Q$ 内の単語の出現回数をカウントし，同じ概念値の単語から出現回数がもっとも多い単語を概念値とペアで「共通概念値表」に保存する．

**手順４**　文書 $P$ と $Q$ 内の単語の概念値が「共通概念値表」にあるならば，それらの単語を「共通概念値表」内の単語で置き換え，新しい文書 $P'$ と $Q'$ を得る．

**手順５**　文字ベースまたは単語ベース n-gram 手法で新しい文書 $P'$ と $Q'$ の文書間類似度 $R$ を式 (1) で計算する．

## 5. 実　　験

提案手法の有効性を検証するために，疑似盗作データを用いた，盗作の定義に応じた６種類の改変への盗作検出に関する実験と６種類の改変を混合させて改変の程度の変化に応じる盗作検出に関する実験を行った．さらに，被験者が指示にしたがって実際に作成したレポートの盗作検出の実験も行った．

### 5.1 改変の種類に関する実験

#### 5.1.1 データ作成

実験に用いるデータは，「ちりめんじゃことしらすの違いについて」というレポート課題を想定し，擬似的に作成した．具体的には，まず，レポート課題から内容語「ちりめんじゃこ」，「しらす」，「違い」をキーワードに Google 検索し，検索結果の上位 20 件の Web ページを取得した．これらのうち，１件目，５件目，１０件目，１５件目，２０件目の Web ページに対し，（不要なタグなどを除去した上で）それぞれ２節に定義した６種類の改変を施し，計３０件の疑似盗作データを作成した．改変の詳細は以下のとおりである．

(1) Web ページからのコピー
改変処理はしない．ただし，実際の盗作検出処理を考え，コピー元のデータはタグなどを自動除去したもので，改変データは人手でタグなどをきれいに除去したものとしている．つまり，両者は完全一致ではない．

(2) 順序変更
(1) で得られたものに対し，ランダムに選んだ約半分の段落に対し順序を入れ替え，さらに各段落内にランダムに選んだ数文の順序を入れ替える[⋆2]．

⋆1 注意：この手法と，「文書 $P$ と $Q$ の単語に対し分類語彙表を用いて概念値を付与し，単語の代わりに概念値の n-gram を形成して式 (1) で文書間の類似度 $R$ を求める．」というような手法とは違う．つまり，「盗作であればごまかすためにその中の数多くの単語は盗作元の単語をそのまま使う代わりにそれらの類義語を使っている可能性が高い」という仮定の元で，提案手法では盗作元の候補文書 $Q$ の単語を用いて盗作 $P$ の単語を置き換えるようにしている．しかし上記手法であれば，盗作 $P$ の単語の置き換えに，盗作元 $Q$ の単語だけでなく盗作 P の単語も用いていることになる．これはつまり，「盗作の中の数多くのオリジナルな単語も盗作の中の他のオリジナルな単語の類義語でごまかしている」という，あり得ない仮定の元での手法となってしまう．明らかにこのような置き換え手法は盗作においては普通採用しないであろうと考える．

⋆2 各 Web ページの文書の長さは数十文しかないものから数百文まで，かなり異なっており，文書内の段落の大きさも一文しかないものから十数文もあるものまで，さまざまであった．さらに，句点や読点の付け方も Web ペー

Vol.2009-NL-194 No.1
2009/11/16

(3) 文削除

(1) で得られたものに対し，文をランダムに選び，約半分を削除する．

(4) 同義語変換

(1) で得られたものに対し，著者らが思いつく限りに，たとえば「違い」を「相違」へと類義語変換を行う．また,「全く」を「まったく」へと，少量ながら，違う表記への変換も行う[⋆1]．

(5) 文末変換

(1) で得られたものに対し，文末表現の「です・ます調」の記述と「だ・である調」の記述を相互に変換する．

(6) 複合改変

(1) で得られたものに対し，上記 (2)~(5) のすべての操作を行う．

#### 5.1.2 予備実験

先行研究では文字ベース n-gram 手法の n を実験せずに 3 に固定していた．本研究では，文字ベースと単語ベースの両手法において，最適な n を選ぶために，予備実験を行った．

文字ベース手法については n を 2 から 5 まで変化させ,(5.1.1 節に述べた) 5 つの Web ページと，それらから得られた 6 種類の改変データとの類似度計算を行った．一方，単語ベース手法については n を 1 から 4 まで変化させ，文字ベース手法と同様の類似度計算を行った．図 1 はこのように得られた類似度の 5 つの Web ページにおける平均値を示す．これらの結果から，n の値が上記範囲で変化させても計算された類似度の改変種類に伴う変化の傾向は似ていることがわかる．また，各改変データの類似度を総合的にみると両手法とも n=2 または 3 あたりが高めの数値になっていることがわかる．一方，実験結果を図示していないが，個々の Web ページについて計算された類似度は，n=2 の場合は文字ベースと単語ベースのどちらにおいても，Web ページによってのばらつきが n=3 の場合より大きかった．Web ページが違っても安定した類似度が得られることが望ましいために，今後の実験では，n-gram の n をすべて 3 に設定した．

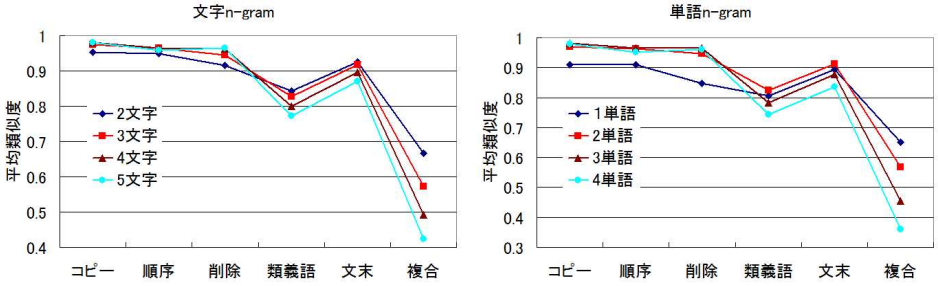


図 1 n を変化させた場合の平均類似度

なお，手法の評価は，盗作に対し高い類似度が得られるかどうかを見るだけでは不十分で，非盗作に対し低い類似度が得られるかどうかを見る必要もある．図 2 は 5.1.1 節に述べた 5 つの Web ページからコピーしたデータと全 20 件の Web ページとの類似度計算の結果を示している．この図から，両手法とも，盗作 (Web ページとそのコピーの間) には 0.9 以上の類似度，非盗作 (Web ページとそれとは別の Web ページからのコピーの間) には 0.2 以下の類似度を出していることがわかる．すなわち，基本的に両手法とも盗作検出に有効であることが言える．さらに細かく見ると，盗作の類似度は両手法ともほぼ同じ値を出しているが，非盗作の類似度は単語ベース手法の方が文字ベースより低く出ていることがわかる．つまり，この結果のみからは，単語ベース手法の方が文字ベース手法より優れていることが言える．

#### 5.1.3 各手法の比較実験

図 3 は各手法の比較実験結果を示す．左側のグラフは，各手法の,(5.1.1 節に述べた) 5 つの Web ページとそれらから得られた各改変データ間の平均類似度を示している．コピーと順序変更についてはどの手法も高い類似度を出している (シソーラス手法の類似度が若干低かった)．削除についてもシソーラス手法を除けばどの手法も高い類似度を出しているが，単語ベース手法の方が文字ベース手法より若干類似度が高かった．一方，類義語変換については，どの手法も類似度が下がっているが，類義語変換に強いシソーラス手法が他の手法よ

---

ジによって千差万別であった．本研究では，より盗作らしい疑似データを作成するために，入れ替えの段落の数と段落内の文の数を一律にするのではなく,「小さな段落であれば，入れ替えの段落の数を多少増やし，段落内の入れ替えの文の数は逆に少なめにする」といったように，人手による各 Web ページの文書の文体などに応じた改変を行った．ただし，このような人手による作成方法では大量なデータを作成することが困難なため，文の数を段落内の文の総数に比例して作成するなど，データ作成の自動化が望ましい．なお，このように作成したデータはすべての手法に用いられるため，入れ替えの文の数などが一律ではないからと言って評価に影響を与えることは基本的にないと思われる．

⋆1 同義語変換については人手に大きく依存したため，改変の程度に Web ページによってばらつきが生じている．ただし,「順序変更」の改変と同様，同じ改変データを各手法に用いているので，手法間の比較評価に関しては公平さは保っていると思われる．

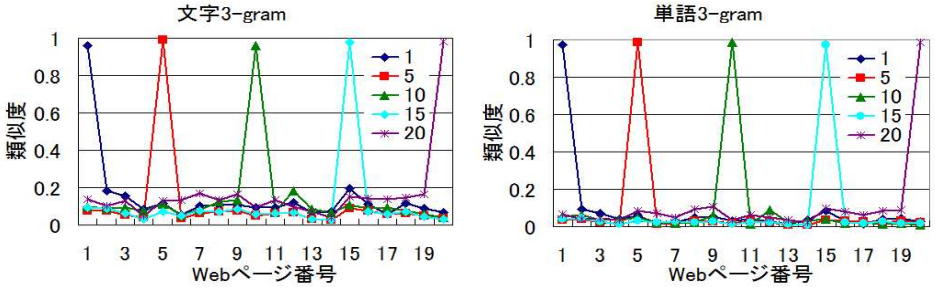


図 2　盗作と盗作元の Web ページまたは他の Web ページ間の類似度

りもっとも高い類似度を出している．また，組み合わせ３文字は文字 3gram より，組み合わせ３単語は単語 3gram よりそれぞれ類似度が高かった．文末変換については文字ベース手法が単語ベース手法より高い類似度を出している．複合改変についてはシソーラス手法を除けばどの手法を用いても類似度が低かった．ただし，元々複合改変に用いたデータは改変の程度が大きいため，もはや盗作でなくなっている可能性が高い．以上の実験結果からは，改変の種類に応じた盗作の検出については，手法間の優劣を判定することができないことが分かった．すなわち，各種類の盗作の検出に各手法は一長一短であると言えよう．

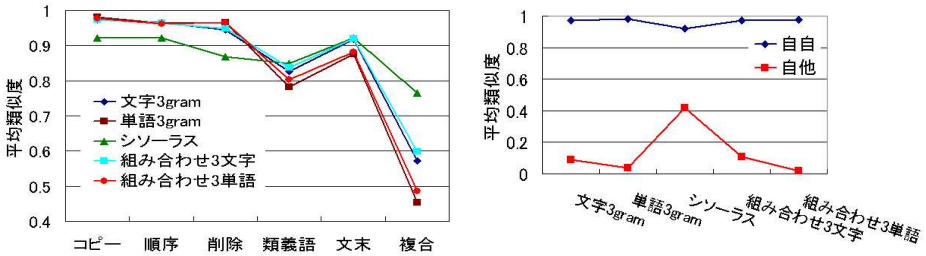


図 3　各手法の比較

一方，右側のグラフは各手法の，５つの Web ページから得られたコピーデータとそれらの Web ページと（グラフ中の「自自」．すなわち「盗作」）の平均類似度と５つの Web ページから得られたコピーデータとその５つ以外の Web ページと（グラフ中の「自他」．すなわち「非盗作」）の平均類似度を示している．単語ベース手法が文字ベース手法より盗作については高い類似度，非盗作については低い類似度を出している．一方，シソーラス手法は非盗作についても相当高い類似度を出している．すなわち，非盗作を盗作として判定してしまう危険性がある．この実験結果からは，シソーラス手法はあまり有効ではないことと，その他の手法では単語ベース手法が文字ベース手法より若干優れていることが分かった．

### 5.2　改変の程度に関する実験

前節の複合改変のように，改変を多く加えると，コピー元との類似度が大きく下がり，盗作かオリジナル作かを見分けるのが難しくなる．本節では改変の程度を変化させ，各手法で計算する類似度がどのように変化するかを調べる．

実験に用いるデータは，前節と同様,「ちりめんじゃことしらすの違いについて」というレポート課題を想定し，擬似的に作成した．具体的には，まず，レポート課題から内容語「ちりめんじゃこ」,「しらす」,「違い」をキーワードに Google で Web 検索し，検索結果の第一位の Web ページを取得した．このデータを元に，以下のような５通りの改変の程度の異なる盗作疑似データを作成した．

（1）　程度１
Web ページの内容を改変せずにこのまま使用する．ただし，前節同様，実際の盗作検出処理を考え，コピー元のデータはタグなどを自動除去したもので，改変データは人手でタグなどをきれいに除去したものとしている．つまり，両者は完全一致ではない．

（2）　程度２
程度１のデータに，順序変更・文章削除・同義語変換・文末変換という四種類の改変を１回ずつ加えて作成した．

（3）　程度３
程度２のデータに上記四種類の改変をさらに１回ずつ加えて作成した．

（4）　程度４
筆者らがオリジナルに作成したデータにコピー元に少し近づくよう多少変更を加えて作成した．

（5）　程度５
筆者らがオリジナルに作成した．

図 4 は上記データを用いた実験結果を示す.「類似度は盗作ほど高くオリジナルほど低い

Vol.2009-NL-194 No.1
2009/11/16

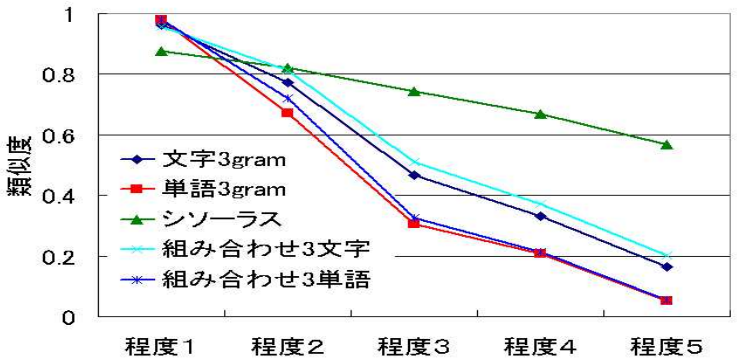


図 4　改変の程度に関する実験結果

方がよい」という評価尺度を用いると，実験結果から各手法の有効性について，組み合わせ 3 単語 ≥ 単語 3gram> 文字 3garm> 組み合わせ 3 文字 ≫ シソーラスという結果が得られた．すなわち，単語ベース手法がもっとも有効で，文字ベース手法も有効であるが単語ベースよりはやや性能が劣った．一方，シソーラス手法は盗作かオリジナルかの判別にあまり有効ではなかった．

### 5.3　被験者実験

前述の二つの実験はいずれも疑似データを用いたものであった．本節は被験者に作成してもらった盗作とオリジナルの両方の課題レポートを用いた評価実験について述べる．

まず，10 人の大学生被験者に「戦争はなぜ起きるのか」という課題レポートを以下のような要領で作成してもらった．

(1)　10 名の被験者の内，6 名は盗作を，4 名はオリジナルなレポートを作成する．
(2)　作成したレポートの長さは 600 字程度である．
(3)　盗作は，「戦争」，「なぜ」，「起きる」をキーワードに Web で Google 検索し，検索結果の上位 10 件までの Web ページのうち３つまでの Web ページの内容をレポートの半分程度に充てて作成する．
(4)　盗作は人にばれないように順序変更・文削除・同義語変換・文末変換という四種類の改変を加えてもよい．
(5)　オリジナルのレポートは，Web 検索はあくまでも参考とし，書籍の情報を参考にしながら自力で作成する．

このように作成した１０人分のレポートのそれぞれに対し，盗作時に利用した上位 10 件の Web ページとの類似度計算を行い，そのうちのもっとも高い類似度を盗作かオリジナルかの判別に使うものとした．実験結果は図 5,6,7 に示す．

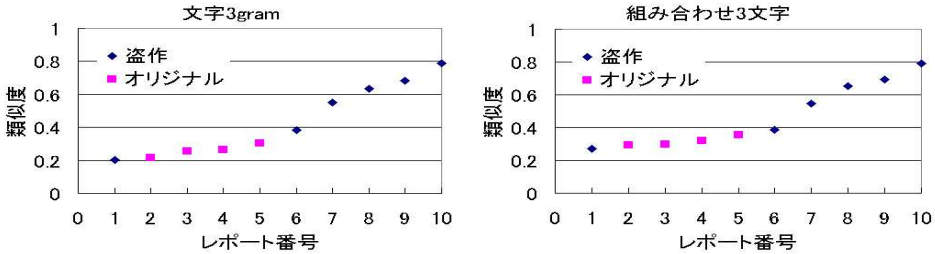


図 5　被験者実験における文字ベース手法の実験結果

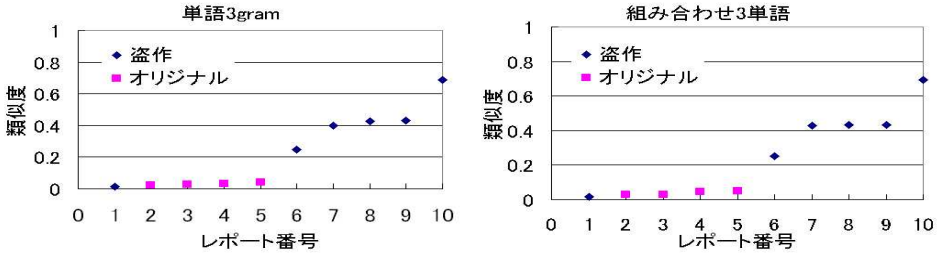


図 6　被験者実験における単語ベース手法の実験結果

これらの結果から以下のことがわかる．まず，図 5,6 に示しているように，文字ベース，単語ベース，そしてそれらとシソーラスとの組み合わせの４つの手法は，１つの盗作を除けば，どれを用いてもオリジナルのものには低い類似度，盗作には高い類似度を出すことができた．類似度がもっとも低かった盗作レポートはその中身を調べると，３分の１ほどは Web からのものであったが，それら Web からのものに対し，細かく文章の入れ替えや

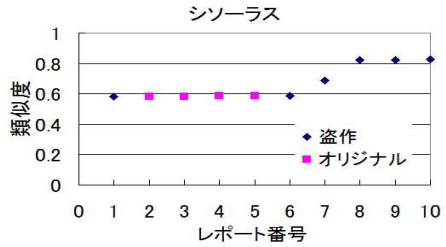


図 7　被験者実験におけるシソーラス手法の実験結果

漢字の変換などを行い，オリジナル文とほぼ変わらないようなレポートになっていたことがわかった．次に，各グラフを細かく見ると，文字 3gram 手法と組み合わせ 3 文字手法はほぼ同じ結果，単語 3gram 手法と組み合わせ 3 単語手法はほぼ同じ結果を出しているが，単語ベースの手法は文字ベースの手法よりよい結果が得られていることがわかる．一方，図 7 に示しているように，シソーラス手法ではオリジナルレポートに対しても 0.6 程度の高い類似度を出してしまった．

## 6. おわりに

Web サイトからの盗作の自動検出システムを構築した．盗作検出に用いる文書類似度計算に，従来の文字ベースの n-gram 手法に加え，新たに単語ベース n-gram 手法，(同義語変換への対処を想定した) シソーラス手法，さらに文字ベースまたは単語ベース n-gram とシソーラスの利用を組み合わせた手法を提案した．これらの手法について，盗作元とされる Web ページに，文末変換・順序変更・文削除・類義語変換によるそれぞれの改変とそれらの混合改変を加えることにより作成した疑似盗作データと被験者が指示通りに作成した課題レポートを用いた類似度の評価実験を行った．その結果，盗作の程度変化の検出や盗作かオリジナルかの区分けには提案手法である単語ベース手法と組み合わせ手法の方が従来の文字ベース手法より優れていることがわかった．

## 参考文献

1) 高橋勇ほか：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， Vol. J90-D, No.11, pp.2989–2999 (2007).
2) 小高知宏ほか：n-gram を用いた学生レポート評価手法の提案，電子情報通信学会論文誌， Vol. J86-D-I, No.9, pp.702–705 (2003).
3) 深谷亮ほか：単語の頻度統計を用いた文章の類似性の定量化–部分的類似性の考慮–，電子情報通信学会論文誌， Vol. J87-D-II, No.2, pp.661–672 (2004).
4) 太田貴久， 増山繁：学生レポート採点支援のためのレポート類似部分発見手法，信学技報, NLC2005-112, pp.37–42 (2006)
5) 国立国語研究所：分類語彙表（増補改訂版），大日本図書 (2004).