様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２５年５月２７日現在

機関番号：１７１０２
研究種目：研究活動スタート支援
研究期間：2012～2012
課題番号：24800049
研究課題名（和文）　テキスト文書のグラフ変換に基づく侵害検知

研究課題名（英文）　Detecting Document Infringement based on Graph Transformation

研究代表者
周　秉慧（CHOU BIN-HUI）
九州大学・システム情報科学研究院・学術研究員
研究者番号：50636793

研究成果の概要（和文）：
近年，インターネットの発展に伴い，文書侵害・盗作行為が大きな問題となっている．文章侵害・盗作を検知する既存の手法は単語系列間の類似度を計算するため，同じ意味をもつが語彙や語順が大いに異なる侵害文章を検知することは困難と考えられている．それを解決するために，本研究は文書を単語間の文法関係を表現できるグラフ構造に変換し，さらにそれに基づく侵害検知手法を開発した．百件の科学論文データを使った実験では提案手法の有用性と有効性を確認した．

研究成果の概要（英文）：
In this research, we tackle the problem of detecting document infringement, which is considered as a severe problem owing to the convenience of Internet. Typical information retrieval methods, stopword-based methods and fingerprinting methods are commonly used to detect infringement by using sequences of words as they appear in the document. As such, they fail to detect infringement when an author reconstructs a source document by re-ordering and re-combining phrases. Because graph structure fits for representing relationships between entities, we propose a novel infringement detection method, in which we use graphs to represent documents by modeling grammatical relationships between words. Experimental results show that our proposed method outperforms two n-gram methods and increases recall values by 10%.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２０１２年度	1,200,000	360,000	1,560,000
年度			
年度			
年度			
年度			
総　計	1,200,000	360,000	1,560,000

研究分野：知能情報学
科研費の分科・細目：情報学・知能情報学
キーワード：侵害検知，文書盗作・剽窃，グラフ変換，グラフマッチング，

科学研究費助成事業（科学研究費補助金）研究成果報告書

１．研究開始当初の背景

近年，インターネットの発展に伴い，科学論文をはじめとする文書の侵害・盗作行為が大きな問題となりつつある．文書の侵害・盗作とは，情報源を明示することなく，他人の文書をそのまま利用するまたはパラフレーズする行為である．

既存の侵害検知手法は情報検索（information retrieval）手法と文書フィンガープリント（fingerprinting）法，ストップワード（stopword）法に大別される．情報検索手法は，共通内容語（common content words）の数や頻度を使ってソース文書と盗作文書の類似度を計算する．文書フィンガープリント法は単語系列のハッシュ値を利用して文書の同一性を判断する．ストップワード法は文書内のストップワードのみを使いソース文書と盗作文書の共通部分を発見する．既存手法のほとんどはソースと盗作文書間の単語系列を比較するため，語彙や語順が大いに異なるパラフレーズによる侵害文書を発見することは困難と考えられる．よって，パラフレーズによる侵害行為を発見する手法が必要となっている．

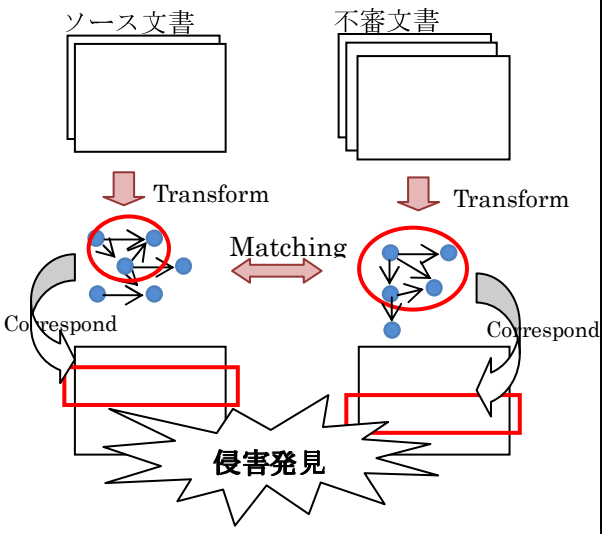

図１　提案手法の概要

２．研究の目的

本研究は，パラフレーズによる文書剽窃・侵害を検知するために，同じ意味をもつが語彙や語順が大いに異なる侵害文書を検出できる新しい検知手法を開発し，計算機システムを利用し手法の適合率や再現率を評価することによってその有用性を確認することを目的とする．本研究において扱う問題は，ソース文書（source document）集合 D_{src} と不審文書（suspicious document）集合 D_{susp} を入力とし，D_{susp} 中の文書 d は盗作文書（plagiarized document）である場合，d と d 中の盗作部分，盗作されたソース文書を出力する問題である．

３．研究の方法

図１は提案手法の概要を示す．まず，ソースと不審文書をグラフに変換する．次にグラフ間の類似する部分構造を侵害と見なし，グラフマッチングを行う．最後に，発見した類似する部分構造に対応する文書の段落や文を出力する．

Source doc

We propose a novel approach for solving the perceptual grouping problem in vision. Our approach aims at extracting the global impression of an image. We treat image segmentation as a graph partitioning problem and propose a novel global criterion, the normalized cut, for segmenting the graph.

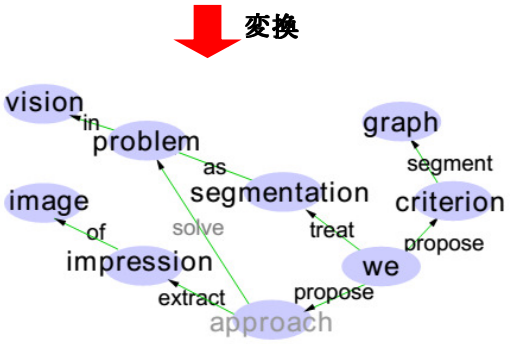

図２　ソース文書をグラフに変換

(1) グラフ変換

同じ意味をもつが語順が異なる文書を検知するために，単語系列による文書表現ではなく，単語間の文法関係を表現できるグラフ構造を考えた．具体的に，文中の名詞をと動詞/前置詞をそれぞれノードとエッジに変換する（図２と３）．それを実現するために，われわれは構文解析器を使って単語間の依存関係（dependency relation）を抽出し，経験的なグラフ変換ルールを提案した．

図３中のテキスト文書は図２の文書から書き換えられたもので，ここで図２と図３の文書をそれぞれソース d_{src} と盗作文書 d_{plag} とする．d_{src} と d_{plag} は語順が異なるが同じ内容を

もつ．共通の単語系列が短いため，従来の手法は図 3 の文書は盗作文書ではないと判断してしまうが，われわれの手法は図 2 と 3 に示すように d_{src} と d_{plag} を類似するグラフに変換する．

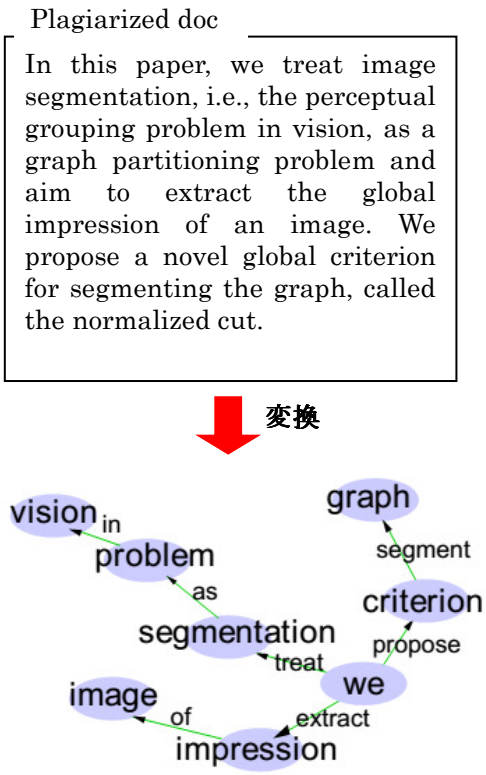

図 3　盗作文書をグラフに変換

(2) グラフの類似部分構造発見

グラフ間の類似する部分グラフ構造を文書間の盗作・侵害と見なし，発見アルゴリズムを開発した．右のアルゴリズムは提案した発見アルゴリズムを示す．入力はソース文書から変換されるグラフ Hと，不審文書から変換されるグラフ G，最少共通ノード数を表すパラメータ k，共通ノード間最短距離を表すパラメータ δ である．出力は G と H間類似する部分グラフ集合である．

アルゴリズムを直観的に説明する．Hにある近い共通ノード（共通ノードを下に説明する）とその経路にあるノードで構成される部分グラフは，Gに類似する部分グラフがあるならば，侵害文書発見になる．アルゴリズムは一つの侵害発見に対し，最大類似部分グラフを発見する（関数 match）．

ノード v と u は互いに同義語または類義語であれば v と u を共通ノードと呼ぶ．アルゴリズムでは，関数 common を用いて共通ノードを発見する．WordNet という概念辞書を利用することにより，異なる語彙を使用した侵害文書を検出することができる．また，自然言語表現の多様性を考慮するため，パラメータ k と δ を用いて部分マッチング（inexact matching）の手法を提案した．

Algorithm of our discovery algorithm

Input: suspicious graph G, source graph H, k, δ
Output: pairs (g, h) of similar subgraphs

```
(Vs, Us) ← common(G, H);
while pop(Vs, Us) do
  (v, u) ← pop(Vs, Us);
  if v and u are not in any pair of (g, h) then
    (g, h) ← match(G, H, v, u, δ);
    if SimNodeNum(g, h) ≥ k then
      output(g, h);
      g ← ∅; h ← ∅;
```

４．研究成果

(1) 実験データと評価指標

DBLP データベースから科学論文を収集し，二種類の人工データ（DBLP_FULL と DBLP_PARTIAL）を用意した．DBLP_FULL と DBLP_PARTIAL データはそれぞれ盗作文書の内容がすべて盗作されたものと内容が一部のみ盗作されたものである．収集した科学論文をソース文書とし，ソース文書に引用される科学論文と人工の盗作文書を不審文書とする．人工の盗作文書を次の方法によって作成する．

- 意味を変更しない前提でソース文書の語順を変える
- ソース文書の単語を任意に選択し同義語または類義語に変更する
- 意味を変更しない前提で新しい語を加える

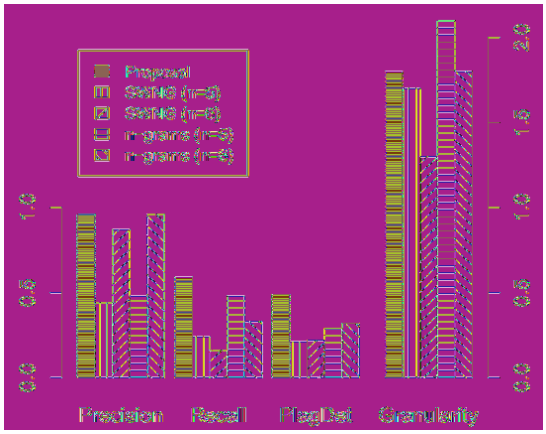

図 4　DBLP_FULL データの実験結果

提案手法を評価するために，検知結果の中にどの程度正解が含まれるかを示す適合率（precision）と，正解のうちどの程度が検知されるかを示す再現率（recall），検知結果はどの程度重複するかを示す粒度（granularity），前三者を結合する指標

PlagDet，四つの評価指標を使用した．PlagDet は次の式で定義される．

$$\mathrm{PlagDet} = \frac{\mathrm{F1}}{\log(1 + \mathrm{granularity})}$$

ただし，$\mathrm{F1} = (\mathrm{precision} * \mathrm{recall})/(\mathrm{precision} + \mathrm{recall})$．

(2) 実験結果

図 4 と 5 はそれぞれ DBLP_FULL と DBLP_PARTIAL データの実験結果を示す．提案手法は二種類の既存手法に比較して，最も高い適合率（DBLP_FULL データは約 95%，DBLP_PARTIAL は 90%）を得た．提案手法は最も低い granularity 値を得ていないが，再現率の値を 10%向上した．同様に，PlagDet 評価指標の結果も提案手法は既存手法より有効であることを示している．

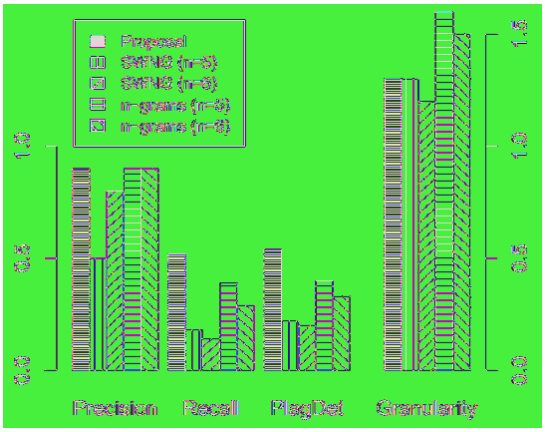

図 5　DBLP_PARTIAL データの実験結果

５．主な発表論文等

〔雑誌論文〕（計１件）

① Bin-Hui Chou, and Einoshin Suzuki, Detecting Academic Plagiarism with Graphs, Extraction et Gestion des Connaissances (EGC), pp. 293-304, 2013.（査読有）

〔学会発表〕（計１件）

① Bin-Hui Chou, Detecting Academic Plagiarism with Graphs, Extraction et Gestion des Connaissances, Toulouse, France, January, 2013.

〔その他〕

ホームページ

http://www.i.kyushu-u.ac.jp/~suzuki/chouj.html

６．研究組織

(1)研究代表者

周　秉慧（CHOU BIN-HUI）

九州大学・システム情報科学研究院・学術研究員

研究者番号：50636793

(2)研究分担者

なし

(3)連携研究者

鈴木　英之進（SUZUKI EINOSHIN）

九州大学・システム情報科学研究院・教授

研究者番号：10251638