様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　４ 年　 ６ 月 １４ 日現在

機関番号：１５３０１
研究種目：基盤研究(C)（一般）
研究期間：2019～2021
課題番号：１９Ｋ１２１３３

研究課題名（和文）文書の画像的処理による効率的な剽窃検知手法の開発

研究課題名（英文）Efficient Plagiarism Detection Based on Image Processing for Documents

研究代表者
馬場　謙介（Baba, Kensuke）

岡山大学・サイバーフィジカル情報応用研究コア・特任教授

研究者番号：７０３８０６８１

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：大規模な文書データに対する効率的な剽窃検知手法を開発した．大量の文書に対して文書間の類似を高速に計算するには，長い計算時間か大きなサイズの検索構造が必要になるという問題点に対し，本研究では画像のフィルタ手法のアイデアを適用することで剽窃検知用データのサイズを削減した．特に，語の近さを表すベクトル表現を用いることによって，字面の単純な一致に基づく剽窃だけでなく，語間の類似を考慮した剽窃の検知について効率化を実現した．提案手法によって実現される文書間類似箇所の検知を，岡山大学の機関リポジトリ内の文書や各部署が持つ研究関連書類に適用し，研究者や研究シーズの検索機能として実装した．

研究成果の学術的意義や社会的意義
本研究の成果により，機械学習技術の大規模文書データへの適用により得られる一般的な知識を，剽窃検知という具体的な応用に利用することができるようになった．機械学習技術によって語を数値ベクトルに変換することができ，これを利用することで文書を画像のように扱うことができる．このアイデアを用いて，画像処理のうち類似する部分を網羅的に調べる手法を文書に適用することができるようになった．結果として，ある程度の曖昧さを考慮した文書間の類似部分の検知を，高速かつ省スペースで行う手法が得られた．

研究成果の概要（英文）：We developed an efficient plagiarism detection method for large-scale document data. Fast computation of similarity over documents for large data requires long computation time or large-size data structure. To solve the problem, we applied the idea of filters for images to documents, to reduce the size of plagiarism detection data. By using a vector representation of words, the proposed method can detect not only plagiarism based on simple string matching, but also plagiarism based on word similarity. We applied the proposed method to documents in the institutional repository of Okayama University and research-related documents owned by each department and implemented it as a search system for researchers and research seeds.

研究分野：情報科学

キーワード：検索　文書解析　剽窃検知　自然言語処理　分散表現

１．研究開始当初の背景

大規模なデータに対する剽窃検知を効率的に行う技術が求められている．インターネットから様々な文書データの入手が可能になり，著作権のある文書や学術論文等からの剽窃が容易になっている．剽窃を文書の類似として定義するには応用ごとの議論が必要であるが，文書間の部分的一致を網羅的に調べることができれば，剽窃を判断する材料として用いることができる．加えて，語句の意味の類似や品詞等の属性の一致を考慮することができれば，言い換え等を含む複雑な剽窃にも対応することができる．しかし，剽窃の元となり得る文書（対象文書）の量が膨大な場合，剽窃の疑いのある文書（問い合わせ文書）の任意の部分について，対象文書の任意の部分との類似を計算するには，長い計算時間か，あるいは大きなデータサイズを要する検索構造が必要になる．これを解決する技術が開発できれば，剽窃の検知とともに抑制の効果が期待でき，人類の健全な創作・学術活動を促すことができる．

２．研究の目的

本研究の目的は，文書間の複雑な類似を高速に計算するための新たなデータ構造を開発することである．文書間の類似として，文書を語の列とみなし，任意の位置ずれについての語の出現順序を保った一致数であるスコアベクトル（図 1）を考える．スコアベクトルは語間の類似度を考慮するよう単純に拡張できる[1]．文書間の類似度として語の出現順序を考慮する場合，各文書のデータ量の削減方法として重要語の抽出やハッシュ化を適用することができない．本研究では，この問題を解決する新たなデータ構造を提案し，その効果の検証を行う．

位置	0	1	2	3	4	5	6	7		
文書A			to	be	or	not	to	be		
文書B	not	to	be							
		not	**to**	**be**						
					…					
						not	**to**	**be**		
							not	to	be	
								not	to	be
類似度	0	**2**	0	0	0	**3**	0	0		

図 1. 文書間のスコアベクトルの例

本研究の独自性は，文書のデータ量削減方法として画像のフィルタ手法のアイデアを用いる点である．画像の画素値等を空間周波数領域に変換し，高周波や低周波成分のみについて逆変換することで，元画像のエッジ抽出や平滑化を行うことができる[2]（図 2）．このとき，空間周波数成分については元画像よりも少ないデータ量を保持することになる．本研究では，このフィルタ手法を文書に適用する．文書間のスコアベクトルは，語のベクトル表現[3]を用いることで行列演算によって求めることができる（図 3）．また，行列間の畳み込みは畳み込み定理と高速フーリエ変換（FFT）によって高速に計算できる[4]ことから，高速な剽窃検知が実現可能である．しかし，文書の行列表現は，語のベクトル表現の次元数に比例して大きくなるため，膨大なサイズになってしまう．本研究で開発を目指す手法では，剽窃検知に求められる精度に応じて，対象文書の行列表現のできるだけ少ない周波数成分を保持する．

図 2. 画像へのフィルタの適用例

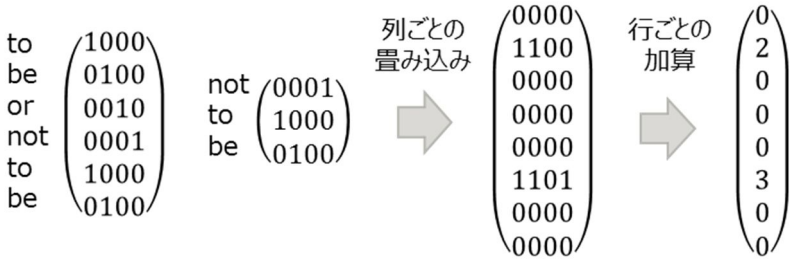

図 3. 行列演算による文書間のスコアベクトルの計算

３．研究の方法

本研究では，語の出現順に加え語間の類似を考慮した剽窃検知についての画像的処理によるデータ削減効果を明らかにした．文書の周波数成分の削減により，剽窃検知精度，計算時間，および剽窃検知用データのサイズの間の効率的なトレードオフが得られるかを検証した．

まず，剽窃検知アルゴリズムの設計を行った．提案技術において語間の類似を考慮するためには，語の近さを表すベクトル表現を用いる必要がある．このため，語のベクトル表現とその獲得手法についてサーベイを行い，本研究で用いるための候補を選定した．

次に，剽窃検知システムの開発を行った．上で得られた文書間類似度算出アルゴリズムと語のベクトル表現を組み合わせて，剽窃検知システムを開発した．

最後に，実験による評価を行った．世界的な剽窃検知コンテストで用いられるデータセットに対し，開発した剽窃検知システムを適用し，剽窃検知の精度，実行時間，検知用データのサイズを測定した．従来技術との比較および組み合わせによる影響を調べ，提案技術の効果を検証した．

４．研究成果

大量の文書に対して文書間の類似を高速に計算するには，長い計算時間か大きなサイズの検索構造が必要になるという問題点に対し，画像のフィルタ手法のアイデアを適用することで剽窃検知用データのサイズを削減した．語の近さを表すベクトル表現を用いることによって，字面の単純な一致に基づく剽窃だけでなく，語間の類似を考慮した剽窃の検知について効率化を実現した．具体的に，本研究では語の出現順に加え語間の類似を考慮した剽窃検知についての画像的処理によるデータ削減効果を明らかにした．語の意味を表すベクトル表現によって，文書はベクトルの列として扱うことができる．ここで周波数成分について特定の領域のみを用いることにより，剽窃検知精度，計算時間，および剽窃検知用データのサイズの間の効率的なトレードオフが得られた．

[1] M. J. Atallah, F. Chyzak, and P. Dumas, A randomized algorithm for approximate string matching, Algorithmica, vol. 29, no. 3, pp. 468-486, 2001
[2] A. K. Jain, Fundamentals of Digital Image Processing, Prentice-Hall Inc, 1989.
[3] C. D. Manning, P. Raghavan, and H. Schuetze, Introduction to Information Retrieval, Cambridge University Press, 2008.
[4] T. H. Cormen et al, Introduction to Algorithms, 2nd edn. McGraw-Hill Higher Education, 2001.

5．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　0件）

1．著者名 Takahiro Baba, Kensuke Baba, and Daisuke Ikeda	4．巻 18(1-3)
2．論文標題 Citation Count Prediction Using Abstracts	5．発行年 2019年
3．雑誌名 Journal of Web Engineering	6．最初と最後の頁 207-228
掲載論文のDOI（デジタルオブジェクト識別子） 10.13052/jwe1540-9589.18136	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計0件

〔図書〕　計0件

〔出願〕　計2件

産業財産権の名称 変化検出プログラム、変化検出装置及び変化検出方法	発明者 馬場謙介	権利者 富士通株式会社
産業財産権の種類、番号 特許、特開2021-179832	出願年 2020年	国内・外国の別 国内

産業財産権の名称 類似文書検索方法、類似文書検索プログラム、類似文書検索装置、索引情報作成方法、索引情報作成プログラムおよび索引情報作成装置	発明者 馬場謙介，野呂智哉，福田茂紀，大倉清司	権利者 富士通株式会社
産業財産権の種類、番号 特許、特願2019-034306	出願年 2019年	国内・外国の別 国内

〔取得〕　計0件

〔その他〕

本研究の成果に基づき具体的なシステムの実装を行った．提案手法によって実現される文書間類似箇所の検知を，岡山大学の機関リポジトリ内の文書や各部署が持つ研究関連書類に適用し，研究者や研究シーズの検索機能として実装した．この検索エンジンを用いた実運用に耐えるシステムは「研究者マッチングシステム」として岡山大学に導入した．

6．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

7．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関