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

平成　３０　年　　６　月　２９　日現在

機関番号：　９２７０７
研究種目：　基盤研究(C)（一般）
研究期間：　2015～2017
課題番号：　１５Ｋ００３１０

研究課題名（和文）複数の指標による類似度を用いた再現率の高い学術論文検索システムの開発

研究課題名（英文）Scholarly Paper Search System Using Multiple Similarity Measures

研究代表者

馬場　謙介（Baba, Kensuke）

株式会社富士通研究所・その他部局等・研究員（移行）

研究者番号：７０３８０６８１

交付決定額（研究期間全体）：（直接経費）　　3,600,000円

研究成果の概要（和文）：学術研究でのサーベイ活動等において，急激に増大する学術論文の内容を効率的に把握することが求められている．本研究の目的は，再現率の高い，つまり，見落としの少ない学術論文検索システムを開発することである．従来の学術論文検索で用いられる指標に加え，本文データの統計的解析による指標を用いることによって，従来の指標では明白ではない関連を持つ論文を発見できる検索システムを開発した．

研究成果の概要（英文）：Researchers are required to grasp the essence of rapidly increasing scholarly papers in their survey activity. The objective of this research is to develop a scholarly paper search system which realizes a high recall, that is, with a small number of oversights. We developed a system which can find papers related to an input paper, using a similarity measure obtained by applying statistical analyses to text data of scholarly papers in addition to usual measures used in existing systems.

研究分野：　情報科学

キーワード：　データマイニング　テキストマイニング　計量書誌学　検索システム　剽窃検知

１．研究開始当初の背景

学術論文誌の種類および掲載論文数は増加の一途をたどっている．専門分野の研究者による研究はもとより，複数の分野を横断する研究や，学生による研究活動の取りかかりを支援するためには，これらの膨大な学術論文の内容を効率的に把握できることが求められる．特に，サーベイ研究や研究の新規性確保のためには，単純な指標では関連が明白ではない学術論文を含めた網羅的な調査が求められる．

２．研究の目的

学術論文間の複数の指標に基づく類似性を利用し，再現率の高い学術論文検索システムを開発することが目的である．学術研究調査では，大量の学術論文を対象にした再現率の高い，つまり，見落としの少ない検索が求められる．従来の学術論文検索システムで用いられる学術論文間の関連の指標に加え，本文データの統計的解析による指標を組み合わせることによって，単純な指標では明白ではない関連による論文を発見することができる．これによって，従来の検索システムを補う新しい検索技術を開発する．

３．研究の方法

研究代表者は，本研究の開始以前に，九州大学の機関リポジトリの高機能化として，引用関係と閲覧状況を考慮した関連論文の検索システムの開発を行っており，この**基本システム**に対する拡張として以下を実施した．

①　学術論文検索システムに対する要求調査

学術論文データベース等の動向や利用状況を調査し，学術論文検索システムに求められる機能を明らかにした．

②　文書の統計的解析による類似度の定式化の調査

語の出現等の統計的解析による文書分類の既存手法について，理論的な実行時間・領域および分類精度の網羅的な調査を行った．①の調査結果と随時比較しながら，学術論文の関連の指標として適した手法の選定を行った．

③　指標の組み合わせによる学術論文間の類似度の定量化

基本システムにおいて既に実装されていた学術論文間の関連の指標と，②で得た本文データの統計的解析に基づく指標を組み合わせ，学術論文間の類似度を数値のベクトルによって表した．

④　学術論文データの収集

基本システムでの検索対象は，当該機関リポジトリに登録されている学術論文である．この対象を，他の機関リポジトリや出版社による電子ジャーナルに拡張する．また，書誌情報のみのデータベースについても，収集の対象とする．具体的には，著作権的な問題を考慮しつつ大規模なデータを収集し格納する作業を行う．

⑤　学術論文検索システムの実装と検証

学術論文検索システムを実装し，①で求められた機能が実現されていることを検証した．

４．研究成果

前述の各研究方法について，それぞれ，以下の成果を得た．

①　学術論文検索システムに対する要求調査

調査の過程で実際の学術論文データに対して解析を行った．特に，論文の引用情報に関する解析で事例としての研究成果を得た．他の論文からの引用数は論文や論文誌，研究者を評価するための指標として用いられるが，引用数そのものに加え研究分野毎の分布や経年変化を解析し，学術論文検索システムに有用な情報が得られることがわかった[発表論文 12]．また，最終的なシステムでの実装へむけて，引用関係の可視化を行った．この成果を論文誌論文として発表した[4]．

②　文書の統計的解析による類似度の定式化の調査

学術論文および引用関係のテンソルによる表現を考えるべきであるとの考えに至った．論文自体のテンソル表現については，語の出現頻度に基づくベクトル化（bag of words，BoW）に加えて，語の次元の少ないベクトル表現（分散表現）に基づくものが新しい技術として注目されていることがわかった．

③　指標の組み合わせによる学術論文間の類似度の定量化

語の低次元ベクトルでの表現（分散表現）を大量の学術論文データから機械学習によって取得し，これをパターン照合的な手法と組み合わせることにより，学術論文間の類似箇所を検出する技術を開発した[9]．類似箇所を剽窃としてとらえ，公開されている剽窃

検知用データセットに対し，検出精度と実行時間についての有用なトレードオフを得た．また，学術論文の引用情報について，研究分野情報との組み合わせによる細分化とその可視化を行った[11]．これらは学術論文間の類似度の詳細な解析を可能にし，この技術の実際の検索システムへの応用方法を示している．また，図書閲覧数の時間変化を利用した利用予測技術を開発した[10]．これは学術論文引用数やダウンロード数に応用可能である．

④　学術論文データの収集

文書の統計的解析により類似度を定式化するために，十分な量のデータを収集するに至った．収集したデータから機械学習により語の分散表現を取得し，これを用いた学術論文中の類似箇所の検索や分類において精度や計算時間の改善が得られた．

⑤　学術論文検索システムの実装と検証

前述の研究成果で得た学術論文間類似度の指標について，実装と検証を行った．学術論文の書誌情報と引用情報から得られる既存手法に加え，語の分散表現を用いた統計的解析による指標を対象とした．申請者の所属研究機関の異動により，所属機関において具体的なシステムを公開することが困難になったため，研究課題全体としては基盤技術の検証に重点を置いた．ここでの研究成果として，語の出現に加えて語の意味を考慮した，学術論文間の類似パターンを高速に検出するシステムを得た[2,3,5]．この語の意味を表現するベクトル表現は，大量の文書データから機械学習的な解析によって得られ，応用先に応じて作成される．つまり，学術論文の本文データから得られた情報（語のベクトル表現によって表される語意）を考慮した類似論文の検索システムを得た．また，副次的な成果として，語のベクトル表現をランダムに決定することによって，単純なパターンの一致の網羅的な検出が，小さな領域によって実行可能であることが分かった[7]．

５．主な発表論文等

〔雑誌論文〕（計 4 件）

[1] Kensuke Baba, Koji Sakaguchi, Mayumi Koyanagi, and Toshiro Minami, A system for paper registration to institutional repositories, IEEJ Transactions on Electrical and Electronic Engineering, vol. 13, no. 5, pp. 763-769, May, 2018.

[2] Kensuke Baba, A Fast Algorithm for Plagiarism Detection in Large-scale Data, Journal of Digital Information Management, vol. 15, no. 6, pp. 331-338, Dec, 2017.

[3] Kensuke Baba, An extension of the FFT-based algorithm for the match-count problem to weighted scores, IEEJ Transactions on Electrical and Electronic Engineering, vol. 12, no. S2, pp. 97-100, Dec, 2017.

[4] 廣川佐千男，伊東栄典，馬場謙介，関連研究探索のための検索可視化システム，情報管理，vol. 58，no. 6，pp. 447-454，Sep，2015.

〔学会発表〕（計 8 件）

[5] Kensuke Baba, Fast Plagiarism Detection Based on Simple Document Similarity, the Twelfth International Conference on Digital Information Management, Sep, 2017.

[6] Toshiro Minami, Yoko Ohura, and Kensuke Baba, Does Student's Diligence to Study Relate to His/her Academic Performance? the Second International Conference on Data Mining and Big Data, Jul, 2017.

[7] Kensuke Baba, Tetsuya Nakatoh, and Toshiro Minami, Vector Representation of Words for Plagiarism Detection Based on String Matching, the 19th International Conference on Human-Computer Interaction, Jul, 2017.

[8] Toshiro Minami, Yoko Ohura, and Kensuke Baba, A Characterization of Student's Viewpoint to Learning and its Application to Learning Assistance Framework, the 9th International Conference on Computer Supported Education, Apr, 2017.

[9] Kensuke Baba, Tetsuya Nakatoh, and Toshiro Minami, Plagiarism detection using document similarity based on distributed representation, the 8th International Conference on Advances in Information Technology, Dec, 2016.

[10] Kensuke Baba, Toshiro Minami, and Tetsuya Nakatoh, Predicting Book Use in University Libraries by Synchronous Obsolescence, the 20th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems, Sep, 2016.

[11] Tetsuya Nakatoh, Hayato Nakanishi, Toshiro Minami, Kensuke Baba, and Sachio Hirokawa, A Visual Citation Search Engine, the 18th International Conference on Human-Computer Interaction, Jul, 2016.

[12] Kensuke Baba, Toshiro Minami, and Eisuke Ito, Modeling Changes in

Demands for Books with Elapsed Time from Publication, the 15th International Conference on Computational Science and Its Applications. Jun, 2015.

〔図書〕（計 0 件）

〔産業財産権〕
○出願状況（計 0 件）
○取得状況（計 0 件）

〔その他〕特に無し

６．研究組織

(1)研究代表者
馬場謙介（富士通研究所）
研究者番号：70380681

(2)研究分担者
南俊朗（九州情報大学）
研究者番号：80315150