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　１４　日現在

機関番号：　１７１０２
研究種目：　基盤研究(C)（一般）
研究期間：　2015～2017
課題番号：　１５Ｋ００４２６

研究課題名（和文）Ｗｅｂからのマイナートピック抽出における信頼性確保に関する研究

研究課題名（英文）Securing reliability in extracting minor topics from the Web

研究代表者
中藤　哲也（Nakatoh, Tetsuya）

九州大学・情報基盤研究開発センター・助教

研究者番号：２０２５３５０２

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：マイナーな情報の抽出に関して(1)Blogに書かれた観光情報(2)学術論文情報のWebサイトから得られる学術論文情報 を分析の対象とした。(1)に関しては観光Blogに含まれる観光資源の候補として抽出された低頻度語に着目し、その背景の分析を行い、2タイプの出現傾向があることを示した。(2)に関しては、研究目的に合う特定の領域からの引用数に限定した新しい評価指標、引用数を用いない評価指標の提案を行い、有効性を示した。また、重要な研究同士を繋ぐ位置にある論文を重要視することで、研究の流れを適切に可視化するシステムを構築した。マイナーな論文を含めた文献の検索、及び分析が行えるようになった。

研究成果の概要（英文）：Regarding extraction of minor information (1) Tourist information written on the blog (2) Academic paper information obtained from the website of academic paper information was analyzed. Regarding (1), attention was focused on the low frequency word extracted as a tourism resource candidate included in the sightseeing blog and its background was analyzed, showing that there are two types of appearance tendencies. As for (2), we showed the effectiveness by proposing a new evaluation index limited to the number of citations from a specific area meeting the research purpose, and an evaluation index not using the cited number. In addition, by emphasizing the papers located at positions connecting important research, we constructed a system that appropriately visualizes the flow of research. It became possible to search and analyze documents including minor articles.

研究分野：　情報工学

キーワード：　マイナートピック　低頻度語　信頼性　学術情報

１．研究開始当初の背景

ブログなどユーザが生成したコンテンツから観光に関する情報を抽出する試みは広く行われているが、観光地名やイベント名自体を求める、あるいはそれらの情報を与えることで関連する情報を得るための研究が主である。これに対して、「何か分からないけど面白い情報」「ごく一部の人にとっては興味深い何か」のような手がかりやキーワードを持たない、そして余り知られていない情報の抽出は、これまで殆ど行われていない。そのような情報の実例として、アニメのシーンに関する実在の場所への観光（アニメ聖地巡礼と呼ばれている）が近年話題となった。そのような体験は、ブログなどに文書として記されていても、ごく一部の有名なもの以外については、特に初期には、関連単語の出現頻度は低く、その他の大量の文書に埋もれることで、その知識を持たない人の目には留まり難い。Web 上に存在していても一般に余り知られていないトピックは、そもそも頻度で評価するのが難しい面がある。それは、従来からの単語の重要度では見つけられないことを意味している。

ここでは、そのような情報のことをマイナートピックと呼ぶことにする。マイナートピックは、（１）それを知っている人はキーワード等を用いることで関連する情報にアクセスする事が可能であるが、（２）それを知らない人は、その情報にアクセスする手段を持たない、という性質がある。

マイナートピックの発見には大きな問題点がある。一つは記載されている情報の頻度が低い事から、それが「意味を持った面白い情報」なのか「単なる偶然による記述」なのか判断が難しい点がある。本研究は、マイナーなトピックに関する情報を持たないユーザに対して、有意義な情報へのアクセスを提供し、発見的な情報収集を可能とすることを目的とする。

２．研究の目的

本研究の目的は主に次の２つである。

(1) 有用なマイナートピックとはどのようなものか？有用性の定義が必要となる。まず実例を調査することで、人間の感覚に合う適切な評価尺度を検討する。必要に応じて、トピックの分野ごとの違いについても検討する。

(2) どのような評価指標を用いれば有用なマイナートピックを抽出できるか？頻度が低いことを前提にしており、一般的に用いられる単語頻度をベースとした指標は使えない。したがって、(1)で検討した評価尺度に従った新しい評価指標が必要である。

３．研究の方法

27 年度はデータの収集と分析を行う。データ収集はブログ、ツイッター、写真の 3 種類についてのクローラーを構築する。分析に関しては、取得済みブログ記事からマイナートピックのサンプルデータを生成し、複数の方法で評価・分析を行う。

28 年度は、27 年度に収集したブログ・ツイート・写真の全体に対して、同じく 27 年度に構築した個々のシステムを適用し、マイナートピックの抽出を試みる。抽出されたトピックを精査し、各システムにおけるマイナートピック抽出の特徴を見極める。更に精度の高い抽出が行えるように各システムの理論的、及び技術的な改良を継続する。その上で有用なシステムはどれであるか、あるいはどのシステムを組み合わせれば最も適切なトピック抽出が行えるかを総合的に判断する。得られた結果を元に、それらを組み合わせたシステムを構築し、評価を行う。更に精度の高い抽出が行えるようにシステム全体の改良を継続する。

29 年度は、28 年度までに構築された個々のシステムを組み合わせ、マイナートピック抽出システムを構築する。抽出の精度等を評価すると同時に、抽出された結果をユーザに適切に表示・提案し、あるいは抽出結果同士や元の文書の関係を可視化する。被験者を用いたシステムの評価を行い、本提案の強みや利点、弱点等を明らかにし、可能であれば、更にシステムの改良を継続する。更に観光情報に関して構築した本システムを他の特定のデータセットで利用するための検討を行い、可能であればそのデータセット専用システムを構築する。更に、汎用システムの構築を検討する。

４．研究成果

マイナーな情報の抽出を評価するために、次の 2 つの情報源を分析の対象とし、それらに含まれる低頻度語に関する分析を行った。(1)ブログに書かれた観光に関する記事、(2)学術論文情報の Web サイトから得られる学術論文の書誌情報。

(1)に関してはこれまでの観光情報抽出に関する研究成果として抽出されたデータに含まれる低頻度語に着目し、ブログ中から観光資源の候補として抽出された低頻度語がどのような背景を持っているかの分析を行ってきた。その結果、2 タイプの出現傾向が存在することが明らかになっている。更に、観光ブログにおける語の出現の頻度とその他の Web 情報における語の出現の頻度の差に関する分析を行った。低頻度語の出現が２タイプの状況から成り立つ事が分かった一方で、それらを分離する手法に関しては確立できなかった。

(2)に関して、初年度は論文を一般的な引用数で評価するのではなく、研究目的に合う特定の領域からの引用に限定した引用数の計上方法を用いることによる新しい評価尺度の提案を行った。この尺度により、引用数が相対的に低い（マイナーな）論文であっても適切に見つけ出す事が可能であることを明らかにした。また、有用な論文を早期に発見

するための手法に関する調査として、引用数変化に関する分析を行い、マイナーな論文が高い評価を受ける過程に関する実例を明らかにした。平成 28 年度は、一般的な引用数と領域限定の引用数の差を用いて、論文の領域特異性を含めた論文の選択や分析を行う事が可能なシステムを構築した。これにより、マイナーな論文を含めて文献の検索、及び分析が容易に行えるようになった。また、論文の引用関係の可視化において、引用数が高くないものの重要な研究同士を繋ぐ位置にある論文を重要視することで、研究の流れを適切に可視化するシステムを構築した。マイナーな論文の一般には把握されていない価値を示すものとなっている。

平成 29 年度は、分析対象をＷｅｂ上から収集した学術論文に限定し、マイナーなものを含む全ての論文に客観的な評価を与えるための手法に関する研究を継続した。学術論文の分析・評価のために、主に次の 3 つの観点からアプローチを行った。（１）Ｗｅｂ上の論文データベースに収集・蓄積されている各論文の情報のうちアブストラクトのみを対象にした分析を行い、後日に高い評価が得られる可能性のある論文を自動分類する。（２）論文が他の論文を引用する際の引用の意味を特定する事で論文間の関係を明らかにし、論文を評価するための補助情報とする。（３）論文の各著者の研究経歴を論文データから分析し、各論文における著者の位置付けを明らかにする。また論文を分析する為の素性として、論文中の特徴語に関して（１）属性選択を用いた重要な特徴語の特定手法、（２）局所的な出現頻度を用いた複合語の重要性の判定手法、に関する研究を進めた。

３年間の研究の結果、主に次の２点に関する研究成果が得られた。（１）観光に関するブログ中に出現する語のうち、低頻度語の出現に関する状況を明らかにした。これはブログなどの文書からの情報抽出に利用可能な研究成果である。（２）Ｗｅｂ上の学術論文に客観的な評価を与える幾つかの手法を提案した。これらにより、マイナーなトピックに関するものを含んだ論文に関してその信頼性を確保するために必要な技術が部分的にであるが構築できた。これらの研究成果により、更なる研究の進展が期待できる状況である。

５．主な発表論文等

〔雑誌論文〕（計　２件）

① Tetsuya Nakatoh, Sachio Hirokawa, Toshiro Minami, Takeshi Nanri, and Miho Funamori: "Quality classification of academic papers using their attributes," Journal of "Artificial Life and Robotics," Volume 23, Issue 2, pp 235—?240, Springer-Verlag New York, Inc., June 2018.

② Tetsuya Nakatoh, Satoru Uchida, Emi Ishita, and Toru Oga: "Performance Comparison on Automated Generation of Coding Rules: A Case Study on ISO 26000," International Journal of Service and Knowledge Management Vol.1, No.1, pp.19—30 (2017).

〔学会発表〕（計２３件）

[1] Tetsuya Nakatoh, Hayato Nakanishi, Sachio Hirokawa: "Journal Impact Factor Revised with Focused View," Proc. of 7th International KES Conference on INTELLIGENT DECISION TECHNOLOGIES (KES-IDT-15), Hilton Sorrento Palace, Italy, 17—19 June 2015. (査読有)

[2] Tetsuya Nakatoh, Hayato Nakanishi, Sachio Hirokawa: "Focused Citation Count: A Combined Measure of Relevancy and Quality," Proc. of 4th International Congress on Advanced Applied Informatics (IIAI-AAI2015), Okayama Convention Center, Okayama, Japan, 12—16 July 2015. (査読有)

[3] Hayato Nakanishi, Tetsuya Nakatoh, and Sachio Hirokawa: "Cause Analysis for Steep Increase of Citation," Proc. of KICSS2015, pp.364—370, 2015. (査読有)

[4] Sachio Hirokawa, Tetsuya Nakatoh, and Hayato Nakanishi: "Accumulated Citation Count as Fertileness of Scientific Article," Proc. of The 2015 International Conference on Computational Science and Computational Intelligence (CSCI'15), pp.119—122, in Las Vegas, USA. DOI: 10.1109/CSCI.2015.74 (査読有)

[5] Tetsuya Nakatoh, Hayato Nakanishi, Toshiro Minami, and Sachio Hirokawa: "Threads and History of Bibliometrics," Proc. of the First CiSAP Workshop in conjunction with ICADL2015, pp.27—31, in Seoul Korea. (査読有)

[6] Tetsuya Nakatoh, Hayato Nakanishi, Kiyota Hashimoto, Toshiro Minami, and Sachio Hirokawa: "Steep Increase Trigger of Citation," Proc. of the 21st International Symposium on Artificial Life and Robotics (AROB2016) (査読有)

[7] Tetsuya Nakatoh, Kiyota Hashimoto, and Sachio Hirokawa: "Analysis of Infrequent Words in Tourism Blogs," Proc. of the 21st International Symposium on Artificial Life and Robotics (AROB2016) (査読有)

[8] Tetsuya Nakatoh, Satoru Uchida, Emi Ishita, and Toru Oga: "Automated Generation of Coding Rules: Text-Mining Approach to ISO 26000," Proc. of the 5th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI 2016), DOI:10.1109/IIAI-AAI.2016.210, pp.154—158, July 10—14, 2016. (査読有)
[9] Tetsuya Nakatoh, Hayato Nakanishi, Toshiro Minami, Kensuke Baba, and Sachio Hirokawa: "Bibliometric Search with Focused Citation Ratios," Proc. of the 5th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI 2016), DOI:10.1109/IIAI- AAI.2016.227, pp.150—153, July 10—14, 2016. (査読有)
[10] Tetsuya Nakatoh, Hayato Nakanishi, Toshiro Minami, Kensuke Baba, and Sachio Hirokawa: "A Visual Citation Search Engine," HIMI 2016, Part I, LNCS 9734, pp.269—274, 2016. DOI: 10.1007/978-3-319-40349-6_17 (査読有)
[11] Kensuke Baba, Toshiro Minami, and Tetsuya Nakatoh: "Predicting Book Use in University Libraries by Synchronous Obsolescence," Proc. of the 20th International Conference on Knowledge-Based and Intelligent Information & Engineering (KES-2016), York, UK, Sept 5—7, 2016. (査読有)
[12] Chao Zeng, Tetsuya Nakatoh, Hiroyuki Takeshita, Ryoji Hisadomi, Masanari Eguchi, and Sachio Hirokawa: "Discovery of Cultural Tourism Preference in Multilingual Tourism Information Site," Proc. of the fifth Asian Conference on Information Systems (ACIS2016), Krabi, Thailand, Oct. 27—29, 2016. (査読有)
[13] Kumiko Kanekawa, Tetsuya Nakatoh, and Sachio Hirokawa: "University Evaluation by Open Data: A Case Study on the Effect of International Exchange Agreement," Proc. of the Eleventh 2016 International Conference on Knowledge, Information and Creativity Support Systems (KICSS2016), Yogyakarta, Indonesia, 10—11 November 2016. (査読有)
[14] Emi Ishita, Tetsuya Nakatoh, Kohei Hatano, and Michiaki Takayama: "An Attempt to Promote Open Data for Digital Humanities in Japanese University Libraries," Proc. of the 18th International Conference on Asia-Pacific Digital Libraries (ICADL2016) LNCS 10075, pp. 269—274, DOI: 10.1007/978-3-319-49304-6_32, University of Tsukuba Japan, December 5—9 2016. (査読有)
[15] Kensuke Baba, Tetsuya Nakatoh, and Toshiro Minami: "Plagiarism Detection Using Document Similarity Based on Distributed Representation," Proc. of the 8th International Conference on Advances in Information Technology (IAIT2016), University of Macau, December 19—22 December 2016. (査読有)
[16] Tetsuya Nakatoh, Sachio Hirokawa, Toshiro Minami, Takeshi Nanri, and Miho Funamori: "Assessing the Significance of Scholarly Articles using their Attributes," Proc. of the 22nd International Symposium on Artificial Life and Robotics (AROB2017), pp.742—746, B-Con PLAZA, Beppu, JAPAN, 19—21 January 2017. (査読有)
[17] Kensuke Baba, Tetsuya Nakatoh, and Toshiro Minami: "Plagiarism Detection Using Score Vectors Weighted by Distributed Representation of Words," HIMI 2017, Part II, LNCS 10274, 2017. DOI: 10.1007/978-3-319-58524-6_34, pp.341—350. (査読有)
[18] Tetsuya Nakatoh, Kenta Nagatani, Toshiro Minami, Sachio Hirokawa, Takeshi Nanri, and Miho Funamori: "Analysis of the Quality of Academic Papers by the Words in Abstracts," HIMI 2017, Part II, LNCS 10274, 2017. DOI: 10.1007/978-3-319-58524-6_34, pp.434—443. (査読有)
[19] Kumiko Kanekawa, Tetsuya Nakatoh, Takahiko Suzuki, and Sachio Hirokawa: "Analyzing Researcher Stage with Last Authorship Ratio: Who is the last author of your paper?," Proc. of 6th International Congress on Advanced Applied Informatics (IIAI-AAI2017) , pp.221—224, (査読有)
[20] Tetsuya Nakatoh, Kenta Nagatani, Kumiko Kanekawa, Takahiro Suzuki, Sachio Hirokawa: "Cluster Analysis of Scientific Citation Context," Proc. of iiWAS2017 will be published by the ACM Digital Library, Salzburg, Austria, 4—6 Dec. 2017. (査読有)
[21] Yasuhiro Yamada, Yuusuke Himeno

and Tetsuya Nakatoh: "Weighting of Noun Phrases Based on Local Frequency of Nouns," Proc. of SCDM-2018, AICS volume 549 by Springer , Johor, Malaysia, 6—7 Feb 2018. **(査読有)**

[22] Tetsuya Nakatoh, Toshiro Minami: "Reducing Computational Effort for Plagiarism Detection with Approximate String Matching," Proc. of SCDM-2018, AICS volume 549 by Springer , Johor, Malaysia, 6—7 Feb 2018. **(査読有)**

[23] Kumiko Kanekawa, Tetsuya Nakatoh, Takahiko Suzuki, and Sachio Hirokawa: "Assessment of Doctoral Supervision of International Students," Proc. of International Conference New Perspectives in Science Education, Firenze Italy, 22—23 March 2018. **(査読有)**

〔図書〕（計　2件）

① Tetsuya Nakatoh and Sachio Hirokawa: "Extraction of Tourism Objects from Blogs," Tourism Informatics: Intelligent Systems Reference Library, Volume 90, 2015, pp.43—58, 15 Jul 2015.

② Sachio Hirokawa, Tetsuya Nakatoh, Hiroto Nakae and Takahiro Suzuki: "Discovery of Implicit Feature Words of Place Name," Tourism Informatics: Intelligent Systems Reference Library, Volume 90, 2015, pp.31—42, 15 Jul 2015.

〔産業財産権〕

○出願状況（計　0件）

○取得状況（計　0件）

〔その他〕

6．研究組織

(1)研究代表者

中藤　哲也（NAKATOH, Tetsuya）
九州大学・情報基盤研究開発センター・助教
研究者番号：20253502

(2)研究分担者

廣川　佐千男（HIROKAWA, Sachio）
九州大学・情報基盤研究開発センター・教授
研究者番号：40126785

池田　大輔（IKEDA, Daisuke）
九州大学・システム情報科学研究研究院・准教授
研究者番号：00294992

山田　泰寛（YAMADA, Yasuhiro）
島根大学・総合理工学研究院・助教
研究者番号：50529609

(3)連携研究者

(4)研究協力者