氏　名	荒井 俊介
学 位 の 種 類	博　士（図書館情報学）
学 位 記 番 号	博 甲 第　８７４３　号
学位授与年月日	平成　３０年　３月　２３日
学位授与の要件	学位規則第４条第１項該当
審 査 研 究 科	図書館情報メディア研究科
学位論文題目	SNS に書かれた疑問記事の自動的な収集・分類方法に関する研究

主	査	筑波大学	教授	博士（図書館情報学）	緑川 信之
副	査	筑波大学	准教授	博士（教育学）	辻 慶太
副	査	筑波大学	教授	博士（工学）	佐藤 哲司
副	査	筑波大学	教授	博士（教育学）	芳鐘 冬樹
副	査	慶應義塾大学	教授	博士（図書館・情報学）	岸田 和明

論　文　の　要　旨（2,000 字程度）

近年，Twitter やブログといった SNS には様々な書き込みが行われ，その中には何らかの疑問が書かれたものがある。本研究はそうした疑問が書かれた記事（以下「疑問記事」）を集めて提示し，回答が付くようにするサイト（以下「本サイト」）を想定し，その構築方法を提案したものである。

第 1 章では本研究の背景と目的，意義について述べている。Twitter やブログ中の疑問記事は各所に分散しているため，それに答えられる者の目に触れることが少ない。そこでそうした疑問記事を集めて提示し，回答を呼び掛けるサイトの有効性を述べている。本研究では特に図書のタイトルに関する疑問記事に焦点を当てている。

第 2 章では関連分野の先行研究を整理し，本研究の位置付けを明らかにしている。本研究はテキストマイニングを一部行う研究とも言えるが，2.1 節ではブログ，Twitter，その他の Web テキストに関するテキストマイニングの先行研究を体系的にまとめ，テキストマイニングを図書館情報学分野に応用した事例も取り上げている。本研究の応用として,図書館員が本サイトを利用するなどしてレファレンスサービスを広く一般に PR することが考えられるが，それを踏まえて 2.2 節ではレファレンスサービスのアウトリーチや Web を利用したデジタルレファレンスサービスの研究，及び質問と回答の対を大規模に蓄積し同じような質問に対応できるようにした国立国会図書館のレファレンス協同データベースの研究をまとめている。2.3 節では本研究の主要な提案の 1 つであるテキスト自動分類に関連した研究，さらには機械学習に関する様々な手法をまとめている。

第 3 章ではブログや Twitter に書かれた疑問を収集・提供する Web サイトの構築手法を提案している。本研究は構築手法を提案するものであり，実際のサイト構築までを行

うものではないが，本研究が考えるサイトの有効性を予備的に示すための実験は行っている。その内容は 3.1 節にまとめられている。まず「本　タイトル　思い出せない」，「本　題名　思い出せない」という 2 組の検索語を用いてサーチエンジンで検索した結果の中には一定数の疑問記事が存在することから，本研究にはニーズがあることを示している。さらに発見した疑問をデジタルレファレンスサービスと Q&A サイトで質問し，両者の回答が高い割合で一致することから，それらの疑問はレファレンスの専門家でなくても回答可能であることを示している。最後に得られた回答を，疑問記事の著者のブログや Twitter に書き込み,回答に対して感謝のコメントが著者から付くことを確認し，知らない者からの回答も歓迎されることを実証している。

こうした予備調査を踏まえ，3.2 節では疑問記事をサーチエンジンに高精度で出力させる検索語の特定方法や，出力されたページ群から疑問記事だけを抽出するテキスト自動分類手法について述べている。3.3 節ではそうして得られた疑問記事を本サイトに表示する上で有用と思われる自動分類方法，具体的には日本十進分類法（NDC）に基づいて分類する手法について述べ，3.4 節では実際に得られたツイートを NDC に基づいて分類する実験方法について述べている。以下ではそれぞれについて述べる。まず 3.2 節では疑問記事に偏って出現する表現の抽出方法を提案している。具体的には MeCab が出力した形態素の 1〜4 個の連続列のうち，疑問記事に偏って出現する列を特定し，サーチエンジンへの入力とすることを提案している。さらにテキスト自動分類では，1〜3 形態素列の出現頻度を特徴ベクトルとし，Support Vector Machine（SVM），Naive Bayes，決定木，ブースティングの 4 つの機械学習手法を用いて，疑問記事と非疑問記事へのテキスト自動分類を行う方法を提案している。3.3 節では 3.2 節の方法で収集された疑問記事を NDC に基づいて分類する手法を示している。具体的にはレファレンス協同データベース中のレファレンス事例を学習用データとし，各事例に挙げられている参考資料，質問文，及び両者を共に用いて特徴ベクトルを作成し，先述の機械学習手法に掛けることを提案している。3.4 節では抽出したツイートを人手で NDC に基づいて分類し，3.3 節の手法を適用して，分類の精度や再現率などを検証することを提案している。

第 4 章では，第 3 章で提案した手法に関する実験結果を示している。まず 4.1 節では 3.2 節の結果を示しており,「タイトル／が／思い出せ／ない」といった 4 形態素列が検索に有効であることを示している。4.2 節では 3.3 節の参考資料，質問文，両者を用いた手法の結果を述べ，質問文だけを用いる有効性などを示している。4.3 節では NDC に基づくツイートの自動分類が高い精度で行えたことを示している。

第 5 章では，これまでの議論をまとめ，本研究が提案した手法の有効性について総括している。さらに今後の課題を提示している。

審　査　の　要　旨（2,000 字以上）

【批評】

第 1 章では，研究の背景として，Twitter やブログといった SNS には様々な書き込みが行われ，その中には何らかの疑問が書かれたものがあること，それらを集め回答を促すサイトを構築すれば社会的に有用であることが述べられている。何らかの疑問を抱いたとしても，それを Yahoo!知恵袋のような Q&A サイトに書き込んだり，図書館のデジタルレファレンスサービスを利用したりすることは一般に心理的ハードルが高い。だが，SNS 上の知人に回答を求めて書き込むことは容易である。本研究はそのような着眼点に基づいて開始され，SNS 上の疑問が書かれた記事（以下「疑問記事」）を集めて提示し，広く回答を促すサイトを構築するという現実的な解を提案している。現在，そのようなサイトは少なくとも日本語に関しては存在せず，想定するサイトの独創性も十分と言える。

第 2 章では，テキストマイニング，テキスト自動分類，機械学習，レファレンスサービスといった本研究に関連する分野の先行研究を体系的に整理している。本研究の性質上，テキストマイニング，テキスト自動分類，機械学習の先行研究を網羅的に整理するのは当然であるが，レファレンスサービスに関する文献も整理し，本研究が想定するサイトをレファレンスサービスのアウトリーチに用いることを提案している点は注目に値する。即ち，図書館員が本サイトを通じて回答し，図書館のレファレンスサービスを PR することで，同サービスの認知度を向上させるという発想は独創的であり，デジタルレファレンスサービスとの連携も予感させるものである。

第 3 章の 3.1 節では，まずブログや Twitter 中の疑問記事を収集・提供する Web サイトを予備的に構築し，その有効性を示している。具体的には，SNS 上には実際に疑問記事が存在し，それらの疑問は Q&A サイトによって一定の回答が得られ，しかもデジタルレファレンスサービスによる回答とほぼ一致するという意味で，レファレンスサービスの専門家でなくても回答可能であり，またその回答を疑問記事を書いた SNS に書き込むと著者から感謝のコメントが得られることを調査実験によって明らかにしている。本研究はサイト構築までは行わず，構築に向けた手法の提案を行うものであるが，本研究が想定するサイトは有用性が高いことを実証的に示した点は評価できよう。

第 3 章の 3.2 節以降と第 4 章では，本研究が提案する手法と実験結果が述べられている。本研究の提案手法は，(1)サーチエンジンで疑問記事候補を高精度でヒットさせ，(2)それら候補の中からテキスト自動分類で疑問記事を抽出し，(3)想定するサイトでのディレクトリ型検索の実現に向けて，記事を日本十進分類法（NDC）に基づいて自動分類する，というものである。(1)では疑問記事と非疑問記事を MeCab によって形態素に分割し，前者に特徴的な 1～3 形態素列を頻度に基づいて特定し，検索に用いることを提案している。この手法自体は特に新奇性のあるものではないが，疑問記事と非疑問記事の大規模なサンプルを独自に構築し，疑問記事に特徴的な形態素列を数量的に明らかにしている点は評価できる。(2)ではこれら 1～3 形態素列の出現頻度を特徴ベクトルとし，Support Vector Machine（SVM），Naive Bayes，決定木，ブースティングという 4 つの機械学習手法を用いて自動分類を行っている。様々な手法を適用した点，及び SVM とブー

スティングが最も効果的であることを明らかにした点は評価できる。(3)では国立国会図書館のレファレンス協同データベースのレファレンス事例を学習用データに用い，そこに挙げられている参考資料，質問文などから計算コストの少ない独自の特徴ベクトルを構築する手法を提案している。構築した特徴ベクトルは再びSVMなどの機械学習への入力としているが，そこではNDCに基づいてテキストを自動分類する先行研究手法との比較も行っている。結果として，本研究手法の方が先行研究手法より分類の精度・再現率等が高いことを示している。既存の手法よりパフォーマンスが高い手法を開発した点は大いに評価できる。ただし，本研究手法に関しては，特徴ベクトルとして単純な頻度を用いているが，それよりもTF-IDFといった偏りを考慮した尺度を利用したり，分散表現を用いた方がパフォーマンスが良くなった可能性がある。また上に挙げた4手法の他に畳み込みニューラルネットワークを用いるという選択肢もあった。これらの検証を行っていない点は本研究の弱点と思われるが，少なくとも先行研究手法より効果的な手法は提案しており，欠点と言うほどではない。

第5章では，本研究の成果と限界を述べている。本論文では自動分類に失敗した事例に関する誤り分析が十分には行われていない。誤り分析を行えば今後の改善方向をより明確に示すことができたと思われる。だが独創的なサイトを提案し，その構築に向けた手法を様々な角度から検証し，先行研究手法よりパフォーマンスが高い手法を提案したという点で，本研究には一定の有用性があると言える。

以上を総合的に判断すると，本論文は図書館情報学の学位論文として十分な内容を有すると認められる。

【最終試験結果】

平成30年1月23日，図書館情報メディア研究科学位論文審査委員会において，本論文について著者に説明を求めた後，関連事項について質疑応答を行った。引き続き，「図書館情報メディア研究科博士後期課程（課程博士）の学位論文審査に関する内規」第23項第3号に基づく最終試験を行い，審議の結果，合格と判定された。

【結論】

よって，本学位論文の著者は博士（図書館情報学）の学位を受けるに十分な資格を有するものと認める。