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　24 年　5 月　24 日現在

機関番号：10101
研究種目：基盤研究(B)
研究期間：2009～2011
課題番号：21300029
研究課題名（和文）　異なる特徴を持つニュースサイトを比較対照する世界ニュース分析システムの研究
研究課題名（英文）　Research on world news analysis system that supports to compare news site with different characteristics
研究代表者
吉岡　真治（YOSHIOKA MASAHARU）
北海道大学・大学院情報科学研究科・准教授
研究者番号：40290879

研究成果の概要（和文）：
本研究は、日中韓の異なる言語で記述されているニュースサイトの内容を比較対照することにより、ニュース分析を行うシステムを提案した。この中で、多様な観点からのニュースを比較するためのマルチファセット分析と、その分析を支えるための、地名などのデータベースの構築、意見分析システムの構築を行った。さらに、このようなシステムを評価するために、システム利用前後の知識の変化に注目したシステムの分析手法の提案を行った。

研究成果の概要（英文）：
In this research, we propose a news analysis system that supports to compare articles from news sites in different countries. This system can support to analyze articles from multi facet (e.g., opinion, named entity). In this research, we propose to construct named entity database and opinion analysis system. In addition, we also propose a framework to evaluate such interactive system by using information that characterizes change of the users' knowledge about task domain.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
平成 21 年度	4,400,000	1,320,000	5,720,000
平成 22 年度	5,500,000	1,650,000	7,150,000
平成 23 年度	3,300,000	990,000	4,290,000
年度			
年度			
総　計	13,200,000	3,960,000	17,160,000

研究分野：　複合領域
科研費の分科・細目：情報学・メディア情報学・データベース
キーワード：情報検索、テキストマイニング、マルチファセット、意見分析、検索システムの評価、地理情報

１．研究開始当初の背景

近年、インターネットにより、様々な国のニュースサイト（新聞や CNN などのニュースチャンネルのサイト）の情報が利用可能になっている。また、Google ニュースのようなニュースアグリゲーションサイトを用いると、最新のニュースをトピック毎に分類して読むことができる。

しかし、これらの既存システムは、多くのニュース記事を集めて分類するだけである。一方、ニュースの価値は、ニュースの情報源と関連する。例えば、外国の新聞における日本の記事は、外国の日本に関する興味を現すと考えられる。そのため、日本と異なる記事の

取り扱いがされているという情報や、通常と異なる報道がされているという情報は、外国からみた日本を理解するために有用な情報となる。
このような多数の新聞記事を分析する代表的な方法としては、トピックによるクラスタリングや、時系列に注目した新規トピックの発見・継続して報道されているトピックへの追従といった研究がなされている。しかし、これらの研究では、上記で指摘した情報源であるニュースサイトの違いなどについて考慮していない。
これに対して我々は、ニュースの分析において情報源の違いを考慮に入れることが重要であると考え、特定のトピックにおける共起語の傾向の違いから、ニュースサイトの特徴を分析する方法の提案や、新聞記事からの意見文抽出などの研究を行っている。また、異なる言語の情報からユーザの情報要求に応じた文書を検索するための言語横断検索や多観点でブラウズと高精度サーチをするファセット型探索的検索システムの提案を行っている。これらの技術を統合することにより、単純な情報の集約とは異なるレベルで世界のニュース情報を分析するシステムが構築可能と考えている。

２．研究の目的

本研究では、国ごとの興味の違いといったニュースサイト毎の特徴を考慮した分析や、意見分析、時系列ごとの分析などの多観点から分析を可能とする下図のような世界ニュース分析システムの構築をその目標とする。
そのために、以下のようなサブトピックについて、研究を行う。

(1) 観点情報の生成
意見分析を行い、異なる意見の比較のための観点を生成するだけでなく、地名や人名を分類・整理し、ニュース記事と対応付けることにより、ニュースに関連する様々な観点を提供する。

(2) 情報源の分析
対照解析やバースト解析などを行うことにより、異なる国の違いを発見するために有用な情報をマイニングする。

(3) 多観点情報の可視化
これらの情報を提示するためのマルチファセット分析のユーザーインターフェースを構築する。

３．研究の方法

本研究期間では、まず、最初に個別の研究者が持つこれまでの研究成果を統合した世界ニュース分析システムのプロトタイプシステムを構築した。
このシステムをユーザに提供し、その結果を分析したところ、様々な情報源の分析手法の性能を議論するためには、これらの手法に提供する一次的な入力データの品質が重要であることが確認された。このため、(1)を中心とした観点情報の生成の研究と、(3)の可視化に関する研究に力点を置き、研究を行った。
また、本システムの評価を行った際に、システムの思想を理解してもらったユーザからは、有意義であるというコメントをいただけたものの、平均的なシステムの有用性を主張することが困難であった。この問題に対処するために、システムの利用前語のユーザの知識量の変化に注目した分析の枠組みを提案した。

４．研究成果

(1) 観点情報の生成

観点情報の生成のための研究として、意見分析システムの研究と、Wikipedia を利用した地名や人名に関する多言語の固有名詞のデータベースの構築を行った。

①意見分析システム

意見分析システムの研究としては、多くの意見分析システムの研究においては、機械学習などにより、特定のドメインに対して性能を発揮する意見分析システムの研究が行われている。しかし、これらのシステムでは、他ドメインの意見文に適用した際に、十分な性能を発揮しない場合があることが確認された。一方、機械学習のためのコーパスをドメインごとに、大量に用意することは、非常にコストがかかるため、既存の学習結果を活用した意見分析システムの開発が求められる。

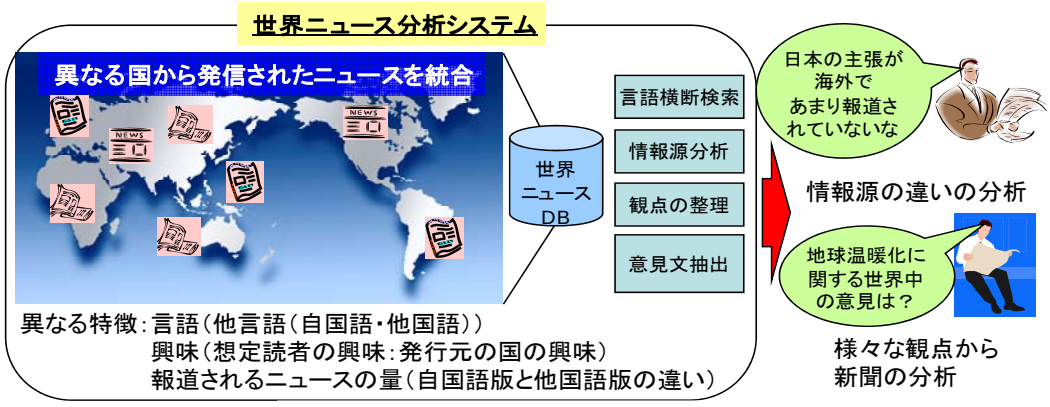

図１：世界ニュース分析システム

本研究では、この問題を解決するために、ドメインごとに学習を行った意見文抽出システムをアンサンブル学習の枠組みで組み合わせることにより、精度をあまり低下させずに、再現率を向上させる方法を提案した（学会発表①）。
具体的には、NICT 意見(評価表現)抽出ツールを基本システムとし、複数のコーパスを用い、複数の意見抽出システムを構築した。
予備的な分析結果から、コーパスが不足している場合には、意見でない文を意見であると誤判定する場合より、意見である文を意見でないと判断する抽出漏れの影響が大きいことが判明した。そのため、アンサンブル学習の枠組みとしては、単純に多数決ではなく、対象ドメインで学習した意見文抽出システムが意見としたものは信頼し、それ以外の文について、他のドメインのシステムが一定数以上意見文と判定したものを意見文とするという枠組みを提案した。
この結果、多くのドメインで、精度をほとんど落とすことなく、再現率が向上し、F 値で判定した場合の性能の向上が確認された。
②固有名詞のデータベースの構築
多言語の新聞記事を比較対照するために、機械翻訳システムを導入した。しかし、機械翻訳システムの性能の問題から、特に、中国語において、新聞記事の分析に重要な役割を果たす新語、固有名詞（オバマ、マケインなどの人名など）の翻訳間違いが多く存在し、結果として、意味のある分析が行えないといった問題が発生した。
このような問題に対して Wikipedia を利用して、日中翻訳辞書を作成した（学会発表⑪）。
具体的には、Wikipedia の言語間リンクを利用した中日翻訳辞書を作成した。本辞書では、日本語においても外来語であるカタカナ表記の語に対して、2666 件の中国語表記に対する日本語表記を獲得した。ここで作成した辞書を言語グリッドにおける辞書連携サービスを利用することにより、機械翻訳システムに組み込んだ。結果として、「オバマ」などの重要な人名の翻訳間違いが減少し、一時情報の性能の向上が確認された。
また、地名のデータベースについては、座標やその包含関係（日本⇒東京）の情報を持つことが、その分析に有用である。このような地理的なデータベースとして、GeoNames(http://www.geonames.org/) が存在するが、日本語に関する情報が不足しており、そのままでは、日本語の分析に利用できない。そのため、Wikipedia のエントリと GeoNames のエントリ間のリンクを発見し、Wikipedia の言語間リンクを用いることにより、日本語の地名のデータベースの構築を行った（雑誌論文①、学会発表⑦）
本手法では、Wikipedia のカテゴリの情報を用いることにより、Wikipedia のエントリに対応する国名や地域名を推定し、対応する GeoNames のエントリを発見した。本手法を用いることにより、これまでに Wikipedia と GeoNames の間に設定されていなかった多くのリンクを発見することができるだけでなく、既存のリンクの一部にエラーが存在することが発見でき、結果として、本手法がこのようなリンクのメンテナンスに有用であることが確認された。
③固有名詞データを利用した質問応答のための情報検索
②で作成した固有名詞のデータベースを用いた質問応答のための情報検索システムの提案を行った。固有名詞に関する質問応答のための情報検索では、対象となる文書中に固有名詞が存在するか否かという情報が、その有用性を判定するのに、大きな役割を果たす。その性質に注目し、固有名詞を考慮したブーリアン検索式を作成し、より適切な文書を検索するシステムの提案を行った（雑誌論文②、学会発表③）。
本システムでは、初期検索式の検索結果の上位である疑似適合文書と、初期検索式を比較することによって、固有名詞や動詞などの表記の違いを考慮したブーリアン検索式を作成する。また、地名については、GeoNames のデータを用いることにより、地名の包含関係を考慮したブーリアン式の検証を行う。
(2) 情報源の分析
情報源の分析手法としては、既存の研究成果であるコントラストセットマイニングの考え方に基づく相関性の変化に注目した解析を行った。具体的には、相関性の大きなキーワードに注目するのではなく、特定のニュースサイトにおけるキーワードと文書群の相関性とそれ以外のニュースサイトにおける相関性の比をとり、その比が大きいもの(そのサイトでは、それなりに注目を浴びているトピックを表すが、他のサイトではあまり述べられていないキーワード)、その比が小さいものを特徴的なキーワードとして抽出する方法を用いた。
また、時系列の分析を行うために、Kleinberg らによって提案されたバースト分析を各国の新聞記事に行い、その結果を並べて表示することで比較を行う枠組みを提案した。バースト分析とは、特定の期間において注目を得たトピック語ならびに注目された期間を分析する手法であり、対象となるデータベースで注目の浴びているトピックを推定することが可能となる。
(3) 多観点情報の可視化
(1)で作成した様々な観点からの記事を分析するために、特定の検索語を含む記事に含まれるデータをいろいろな観点(ファセット)から可視化するマルチファセット分析シス

テムを作成した。本システムでは、キーワード・人名・組織名・地名・賛否の５つの観点が利用可能である。
また、観点ごとのデータは、時系列グラフ・円グラフ・棒グラフ・(重みつき)頻度の表の形で表現できる。
ユーザは、複数存在するグラフ表示領域毎に、表示させたい観点、表示領域に対して与える絞り込み条件や種類を設定することが可能であり、以下のような分析を支援する。

- 複数の観点による検索条件の設定
  複数の観点を組み合わせた検索条件を設定すると共に、その内容を可視化
  する。
- 複数の国の比較
  比較したい国に対応する形で複数のグラフを表示する。表示内容(例えば、人名)などは､全体で固定し､各グラフに対して、国名を絞り込み条件として与えると､国ごとの違いを並べて見ることが可能となる。

(4) NSContast システムの構築
本研究で作成した NSContrast は、次の機能により、複数のニュースサイトから集められた新聞記事の検索・分析を支援する。

- 国別比較
  検索キーワードを含む記事数の遷移をグラフで表示する。また、その単語が注目を浴びているかどうかについて､その期間についても表示する。
- 記事検索
  検索キーワードだけでなく､ニュースサイトの名前や国などの細かな条件で記事を検索する。
- 国別注目単語比較
  バースト解析により、特定の日に、注目を浴びているキーワードを表示する。
- 今日の話題
  特定の日の記事を似ているものどうしをまとめて表示する｡特定の話題に関する記事をまとめてみることが可能。
- 多観点分析
  特定のキーワードを含む記事のデータを詳細に分析する｡記事中に含まれている人名や組織名の数やその時間遷移などをグラフで表示することが可能。
- 関連語の関係性分析（国別比較）
  特定の条件を満たす記事からその記事によく出てくる特徴的なキーワード（人名・組織名）などを抽出し、その関係をグラフで表現する。
- 関連語の時間遷移の分析
  特定のキーワードを含む記事で､その記事と特徴的に出てくるキーワードを時系列順に表示する。

(5) 評価実験
本システムの有効性を検証するために、首都圏の大学生を対象とした被験者実験を行った（雑誌論文③)。利用したニュースサイトのデータベースは、日中韓米のニュースサイト日本：朝日新聞・日経新聞・読売新聞、米国：CNN、韓国：朝鮮日報・中央日報、中国：人民網)に加え、それぞれの国の言語のニュースサイト(韓国：朝鮮日報、中国：新華社、米国：CNN, New York Times)から獲得した 1 年分の新聞記事データを利用して構築した。
この実験では、より詳細な検索活動とその結果を分析するために、自発的な回答を行うアンケートに加え、課題前後の知識の変化を分析するために、コンセプトマップを用いた情報探索活動の分析を行った。
事前知識によるコンセプトマップの作成
ユーザは各課題について、事前に知っている情報をコンセプトマップの形式で記述する。

A) 課題に関する検索
  ユーザは 2 グループに分かれ､1 グループは NSContrast を利用した分析を行い、もう 1 グループは Google News などの既存のニュースサイトを利用する。また、1 課題ごとに利用するシステムを交代する。
B) 事後知識によるコンセプトマップの作成
  ユーザは検索活動の結果として得られた情報をもとに､新しくコンセプトマップを作成する。
C) アンケートへの回答
  ユーザは､各課題について得られた情報について、以下のアンケートに回答する。
D) 全ての検索後のグラフを作成した後に､検索前のグラフ中にあるノードと検索後のグラフにあるノードとの対応関係があるものについては､赤い丸をつけるという作業を行う。

この実験のアンケートの分析結果からは、NSContrast が Google News などに比べて、有用であるという結論は得られなかった。特に、既に、トピックに関する内容がまとめて整理されているサイトなどを見つけた場合に、ユーザの満足度が高かった。これは、一次情報のみを扱う本システムと比較するのがやや不適切だとも考えられるが、現実問題としては、このようなサイトが存在することは多くあるため、システムの有用性という意味では、より詳細な分析が必要である。
また､NSContrast に不満を持つユーザからは、機械翻訳の質や、記事数の不足といった情報量に関するコメントも多く寄せられた。これは、今回提案しているシステムの問題というよりは、現在、利用可能な要素技術などに起因する問題であり、その影響は、差し引いて評価結果を分析する必要がある。
一方、本システムを利用したユーザからは、各国ごとの興味の違いに関する言及を見つけることができ、所定の目的を果たしていることは確認された。
また、本システムの利用結果について満足度

が高かったユーザの描いた検索前の検索活動前後のコンセプトマップの例を示す。システムへの習熟度も考慮して、最後の３つの課題について作成したコンセプトマップを図 2 に示す。真ん中にある二重丸のノードが検索課題全体を表すノードであり、ユーザは、自分がその検索課題について知っている概念をノードとし、その関係をリンクとするグラフを作成する。

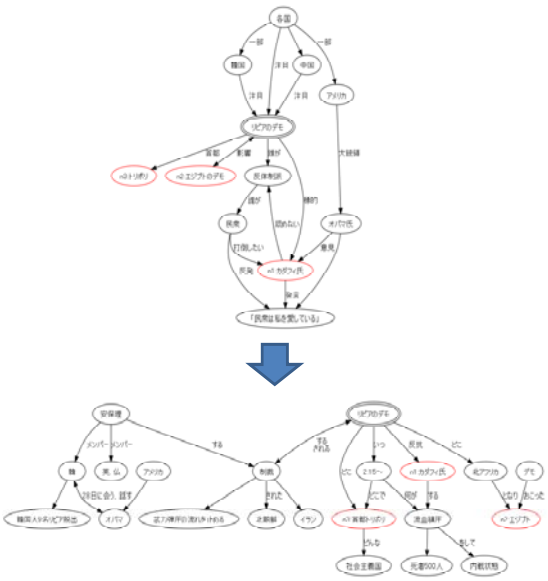

図 2：検索前後の知識の変化（NSContrast）

例えば、図 2 の検索前のグラフからは、ユーザが「リビアのデモ」について、「カダフィ氏」と「オバマ氏」の関係や、「エジプトのデモ」と関係があること、「トリポリ」が主都であることなどを知っていることがわかる。また、「カダフィ氏」、「エジプト」、「トリポリ」については、二つのグラフに対応関係があると判断されたことが示されている。この赤い丸のノードにつながっている概念の変化に注目することにより、ユーザがどのようなタイプの知識を得たかを分析することが可能となる。

この分析の結果、NSContrast を使った場合には、検索前に全く存在しなかったノードが直接トピックのノードと接続している場合が多く、これは、NSContrast を使うことにより、検索前には、全く関連知識を持っていなかった知識に気づくことができ、結果として、幅広い分析が可能になったことを示唆していると考えている。

このコンセプトマップを用いた知識の変化を表現するための枠組みは、自発的な回答によるアンケートでは分かりにくい、知識の変化の質を分析することが可能であることを示唆していると考えている。

(6) まとめ

本研究で提案した NSContrast は、システムの思想を理解していただいたユーザからは一定の評価を得たものの、自発的なアンケート分析の結果からは、全体としてのシステムの有用性が確認できないという状況になった。しかし、この状況を詳細に分析すると、本システムを使うことにより、これまでにはあまり気付かなかったような観点を提示できる可能性があることが確認されている。今後は、この検索前後の知識の変化に注目した分析を、NSContrast のようなインタラクティブな情報探索支援システムの評価を行う手法として確立させると共に、提案システムの詳細な評価につなげていきたい。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　7 件）

① 吉岡真治, 劉亦, 神門典子: Wikipedia カテゴリを用いた Wikipedia と GeoNames 間のリンク発見とメンテナンス. 情報処理学会論文誌データベース（TOD）, Vol. 5, No. 3.（採録決定）

② 吉岡真治: 質問応答のための情報検索への応用を目的とした確率型検索モデルとブーリアン検索モデルの組み合わせ. 情報処理学会論文誌, Vol. 52, pp. 3423-3434, 2011.

③ Masaharu Yoshioka, Noriko Kando and Yohei Seki: Evaluation of Interactive Information Access System using Concept Map. In Proceedings of the 4th International Workshop on Evaluating Information Access (EVIA), A Satellite Workshop of NTCIR-9, December 6, 2011 Tokyo Japan, pp. 20-23, National Institute of Informatics, 2011.

④ Fredric Gey, Ray R. Larson, Jorge Machado and Masaharu Yoshioka: A Micro-analysis of Topic Variation for a Geotemporal Query. In Proceedings of the 4th International Workshop on Evaluating Information Access (EVIA), A Satellite Workshop of NTCIR-9, December 6, 2011 Tokyo Japan, pp. 9-13, National Institute of Informatics, 2011.

⑤ Masaharu Yoshioka: On a Combination of Probabilistic and Boolean IR Models for Question Answering. In Information Retrieval Technology 6th Asia Information Retrieval Symposium, AIRS 2010, Taipei, Taiwan, December 2010 Proceedings, pp. 588-598, LNCS6458, 2010.

⑥ Masaharu Yoshioka: NSContrast: Characterizing the Difference of Interest among Multiple News Sites. In INTERACT 2009 Workshop on Culture and Technologies for Social Interaction,

http://sites.google.com/site/crossculturework shopinteract09/position-papers-1/10-Yoshioka.pdf?attredirects=0&d=1, 2009.

⑦ Masaharu Yoshioka: NSContrast: An Exploratory News Article Analysis System that Characterizes the Differences between News Sites. In SIGIR2009 Workshop on Information Access in a Multilingual World, pp. 25-29, 2009.

〔学会発表〕(計 12 件)

① 高村慎太郎, 吉岡真治, 関洋平: 複数ドメインの意見分析コーパスを用いたアンサンブル学習による意見分析システムの提案. 言語処理学会第 18 回年次大会発表論文集, pp. 235-238, 2012.

② 酒井哲也, 上保秀夫, 神門典子, 加藤恒昭, 相澤彰子, 秋葉友良, 後藤功雄, 木村文則, 三田村照子, 西崎博光, 嶋秀樹, 吉岡真治, Shlomo Geva, Ling-Xiang Tang, Andrew Trotman, Yue Xu: NTCIR-9 総括と今後の展望. 情報処理学会情報基礎とアクセス技術研究会, 2011-IFAT-106, 2011-IFAT-106-5, 2012.

③ Masaharu Yoshioka: ABRIR at NTCIR-9 GeoTime Task Usage of Wikipedia and GeoNames for Handling Named Entity Information. In Proceedings of the 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Quesiton Answering, And Cross-Lingual Information Access, pp. 75-81, National Institute of Informatics, 2011.

④ Fredric Gey, Ray R. Larson, Jorge Machado and Masaharu Yoshioka: NTCIR9-GeoTime Overview - Evaluating Geographic and Temporal Search: Round 2. In Proceedings of the 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Quesiton Answering, And Cross-Lingual Information Access, pp. 9-17, National Institute of Informatics, 2011.

⑤ Yiqi Liu and Masaharu Yoshioka: Construction of large geographical database by merging Wikipedia's Geo-entities and GeoNames. 人工知能学会第 25 回セマンティックウェブとオントロジー研究会, SIG-SWO-A1102-03, 2011.

⑥ 吉岡 真治, 神門典子, 関洋平: 複数国の新聞サイトを比較分析する NSContrast の実験的分析. 情報処理学会情報基礎とアクセス技術研究会, 2011-IFAT-103, 2011-IFAT-103-2, 2011.

⑦ 竹中 均, 吉岡 真治: Wikipedia を用いた地名の包含関係情報の抽出. 2011 年度人工知能学会全国大会(第 25 回)論文集, CD-ROM 2J3-NFC2-2, 2011.

⑧ 吉岡 真治: 質問応答のための情報検索への応用を目的とした確率型検索モデルとブーリアン検索モデルの組み合わせ. 情報アクセスシンポジウム 2010, 2010.

⑨ Masaharu Yoshioka: On a Combination of Probabilistic and Boolean IR Models for GeoTime Task. In Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Quesiton Answering, And Cross-Lingual Information Access, pp. 154-158, National Institute of Informatics, 2010.

⑩ 吉岡 真治: 複数国のニュース比較分析システム : NSContrast. 情報処理学会創立 50 周年記念(第 72 回)全国大会講演論文集, pp. 5-301--5-302, 2010.

⑪ 吉岡 真治: 多言語ニュースの対照分析のための Wikipedia 活用手法の研究. 2009 年度人工知能学会全国大会(第 23 回)論文集, CD-ROM 2G1-NFC5-8, 2009.

⑫ 吉岡 真治: 複数国ニュースサイトの比較対象分析システム NSContrast における評価手法に関する検討. 情報処理学会デジタルドキュメント研究会, 2009-DD-71, 2009-DD-71-1, 2009.

〔図書〕(計 0 件)

〔産業財産権〕

○出願状況 (計 0 件)

なし

○取得状況 (計 0 件)

なし

〔その他〕

ホームページ等

６. 研究組織

(1)研究代表者

吉岡 真治 (YOSHIOKA Masaharu)

北海道大学大学院情報科学研究科・准教授

研究者番号 : 40290879

(2)研究分担者

神門 典子 (KANDO Noriko)

国立情報学研究所・情報社会相関研究系・教授

研究者番号 : 80270445

関 洋平 (SEKI Yohei)

筑波大学・図書館情報メディア系・助教

研究者番号 : 00348468

(3)連携研究者

なし