様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

# 科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ６ 月 ２０ 日現在

機関番号：６２６１５
研究種目：基盤研究(B)
研究期間：2011～2014
課題番号：２３３０００４０

研究課題名（和文）確率的生成モデルの合成による情報アライメントの研究

研究課題名（英文）A Study on Information Alignment by Composite Generative Model

研究代表者
高須　淳宏（TAKASU, Atsuhiro）

国立情報学研究所・コンテンツ科学研究系・教授

研究者番号：９０２１６６４８

交付決定額（研究期間全体）：（直接経費）　　15,300,000円

研究成果の概要（和文）：本研究は潜在トピックモデルを用いた情報の多様な分析法の構築を目的とし、情報に付与された時間や情報間の関連性も考慮した分析モデルを構築した。まず、時間情報とテキスト情報を同時に用いるために、文書にタイプスタンプを付与し、テキストとタイムスタンプを同時に生成するトピックモデルを考案した。さらに論文の引用のように相互にリンクされた文書を生成するトピックモデルに拡張した。本研究の応用として、研究者推薦システムを試作し、多様な情報を活用することにより、共同研究者の推薦精度の向上をはかれることを実験的に示した。

研究成果の概要（英文）：The purpose of this study is to develop topic models for analyzing information in various aspects. We first develop a topic model for handling time as well as text, where we add timestamps to each document. The model generates both text and timestamps simultaneously. Next we extend the model to treat networked documents where documents are linked each other like citations of academic papers. We apply the models to researcher recommendation systems and empirically show that features extracted by the models are effective for recommendation.

研究分野：情報工学

キーワード：トピックモデル　情報推薦　機械学習

１．研究開始当初の背景

大量の情報がネットワーク上の複数のコンピュータに重複して蓄積されている現在、これらの情報を統合する技術が重要になっている。このような問題に対処する技術として情報統合の研究が行われてきた。たとえばGoogle ニュースのように複数の新聞社から発信されるニュースのトピックごとのクラスタリング、ブックマークの共有、遺伝子の類似箇所の発見、論文や文書のコピー検知など情報解析技術を必要とする問題は多岐にわたる。情報解析では、これまで文字列の類似度など表層的類似度に基づいた解析技術の研究が多く行われてきた。また、テキスト情報の解析ではテキスト中に含まれる用語の多義性により表層的な類似度を用いた情報解析では不十分で、テキスト分類などを用いた深層情報に基づく解析も試みられている。しかし、テキスト情報の解析では問題に応じて表層情報と深層情報を組合せて用いることが重要である。

深層情報の抽出に関しては、Blei らのLatent Dirichlet Allocation(LDA)の提案以来、潜在トピックを用いた文書生成モデルの研究が活発に行われている。これらの研究では主にテキストを用語の集合(Bag of Words: BOW)とみなし、用語集合を生成するモデルの獲得に焦点があてられてきた。これらの研究成果は情報解析のための深層情報の抽出技術としても有効であるが、情報解析では関連情報も加味した処理が求められる。たとえば書誌に含まれる同姓同名の著者の識別では、共著関係やタイトル・抄録から得られる研究トピックが有力な手掛かりとなる。テキストを BOW として扱うだけではこのような関連情報を得ることはできず、テキストを要素とするテーブルやネットワークで表される構造をもった情報として扱う必要がある。

時間情報は様々な情報を解析する上でもまた解析された情報を活用する上でも重要である。たとえば、学術論文で扱われる研究トピックにはトレンドがあり、特定の研究トピックは特定の期間に集中的に研究されることが多い。研究トピックに基づいた学術論文の解析では、トレンドを考慮することで情報解析の精度を向上させる試みがなされている。一方、解析された情報を活用する観点からは、最新の研究トピックや最新のニューストピックを注視する上で時間情報は不可欠となる。

２．研究の目的

これまでの情報解析では主にオブジェクトレベルで類似度が計算されたが、前述のような構造情報や時間情報を考慮した情報解析においては、単にオブジェクトの類似度を測るだけでは不十分で、構造をもったオブジェクトの要素レベルの対応が必要になる。たとえば文字列や木構造データの編集距離を用いたマッチングでは、オブジェクト間のアライメントが行われ、対応する文字や木のノードが距離計算の過程で求められる。アライメントによって、類似オブジェクトの対応部分を提示することが可能になり、利用者に対してより詳しい統合情報を提示できる。本研究では、編集距離等で行われる表層レベルでの情報分析を潜在トピック等の深層レベルでの分析に広げるため、多様な統計モデルを開発し、テキストの構造情報および時間情報を考慮した情報解析法を構築することを目的とする。

３．研究の方法

本研究では、潜在トピックモデルを用いた情報の多様な分析モデルの構築とそのモデルの情報推薦システムへの応用に取り組む。

（１）潜在トピックモデルを用いた文書分析

データ構造や分析の目的に応じて複合的な確率的生成モデルを開発する。本研究では、特に時間情報を考慮したモデルおよび文書ネットワークに対するモデルの構築に焦点をあてる。

（２）情報推薦システム

トピックモデルを用いた利用者およびアイテムの潜在情報の抽出と抽出された情報を活用した情報推薦法の研究を行う。情報推薦においては、利用者がアイテムに付与した評価値を用いることができるため、利用者やアイテムの属性に加え、評価値を取り込んだトピックモデルの検討を行う。応用領域として学術情報をとりあげ、学術論文や研究者などの学術情報の推薦システムを試作する。

４．研究成果

(1) 潜在トピックモデルを用いた文書分析

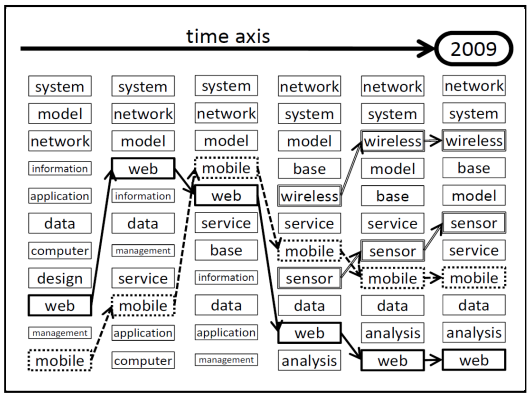
潜在トピックモデルを用いた文書分析に関しては、以下の 2 点が主な成果である。①単語多項分布のディリクレ事前分布を文書のタイムスタンプに依存させることで、時間情報を反映したトピック抽出を実現した[6]。②学術論文の引用関係を文書ごとのトピック確率分布の遷移関係として遷移行列を使って明示的にモデル化することにより、潜在トピックの発展(evolution)の抽出を実現した[3]。以下、それぞれについて解説する。

①文書の時間情報を反映したトピック抽出

トピックモデルに文書の時間情報を反映させる方法としては、Wang 等が提案したTopics over Time のようにタイムスタンプ・データを別の確率分布（ベータ分布）によって生成させる提案もあった。しかし、トピックを表す単語多項分布と時間情報とが別々の確率分布で表現されているため、それらの

間の関連性が明確に得られない問題点があった。

そこで、本研究[6]では単語多項分布の事前分布であるディリクレ分布を文書のタイムスタンプに依存させることにより、タイムスタンプ毎に異なるディリクレ事後分布が得られるようにした。具体的には、トピック数を K、タイムスタンプ数を T とすると、提案のモデルでは、K×T 種類の異なるディリクレ事後分布が得られる。これらの事後分布の推定結果をもとにすれば、同じひとつのトピックについても、異なる時点ごとに異なる単語確率分布が得られることになり、トピックの時間的推移が分析結果として得られる。下に DBLP の論文タイトル・データを分析した結果の例を示す。




この図では、特定のトピック内で、確率が大きい上位 10 単語が 2004 年から 2009 年の範囲でどのように変化するかを表している。各年で異なるディリクレ事後分布を得ることができているため、このように時間情報とトピックの関連性が明確に分析される。

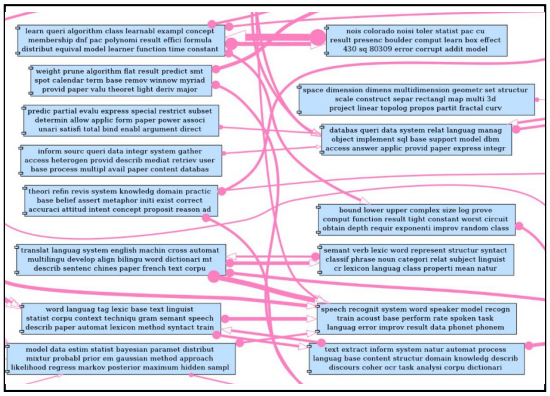
なお、この研究はその後も継続しておこなわれ、文書の時間情報を利用したさらに新しい手法の提案にもつながっており、意義のある成果だったと言える。

② トピックモデルにおける文書間引用関係の明示的モデリング

与えられた文書集合に潜在するトピック間の関連性をモデル化した研究としては Nallapati 等の研究がある。しかし、この提案手法にはパラメータ数が多いという問題点があった。なぜなら、提案されている確率モデルでは個々の引用関係（ある論文が別の論文を引用したという個々の関係）についてトピック遷移を表すパラメータを別々に設定しており、引用の総数を M とすると、通常のトピックモデルに比べて追加されたパラメータ数が M×K 個にも及ぶためである。

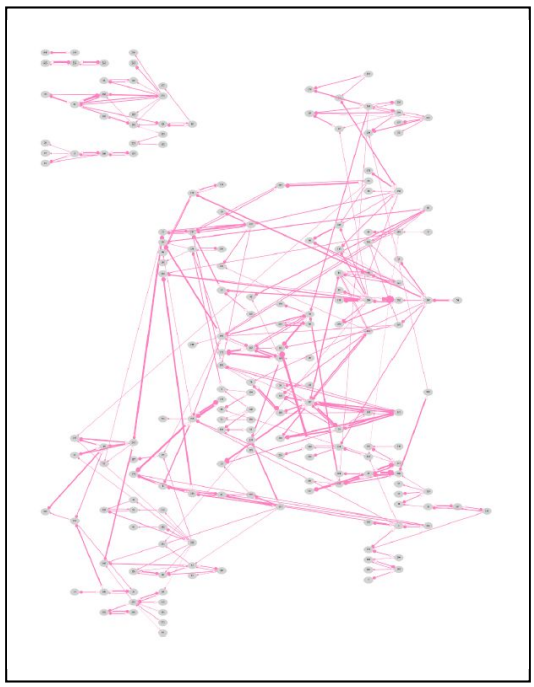
その一方、今回の提案[3]では、論文間の引用関係を、二つの論文におけるトピック確率分布の遷移関係として明示的にモデル化した。そして、すべての引用関係におけるトピック確率分布の遷移を、同じひとつの遷移行列で表している。よって、追加のパラメータ数をトピック数の 2 乗におさえることができた。

推定計算はやや複雑になるが、変分ベイズ推定における evidence の lower bound の評価に工夫を加えることで、計算が出来るかぎり煩雑にならないようにしている。さらに、トピック数の 2 乗に比例する計算量に対処するため、NVIDIA CUDA 向けの実装を C 言語でおこない、グラフィックカードを利用した推定計算の高速化も実現している。




このように、本研究の特徴は主に推定計算やその実装での工夫にあるが、得られた分析結果の可視化も試みている。上に例を示す。

青いボックスがトピックに対応しており、生成確率の上位 21 語がそれぞれのボックスに書かれている。ピンク色の有向枝が各トピックから別のトピックへの遷移を表しており、枝が太く描かれているほど遷移確率が高いこと意味している。この図はすべてのトピック間の遷移関係を表した図の一部であり、実際には数百のトピックの間の遷移が人目で見渡せるような可視化が得られている。トピック間の遷移の全体を俯瞰した図を下に示す。



本研究ではトピック間の遷移を遷移行列によって明示的に記述することで興味深い

分析結果を得ることができた。しかし、推定計算にやや大きな近似を入れていることや、ランダムな初期化からスタートする推定結果の再現性が弱いことなど問題点もあり、さらなる改良の余地があると思われる。

(2) 潜在トピックモデルに基づいた情報推薦システムへの応用

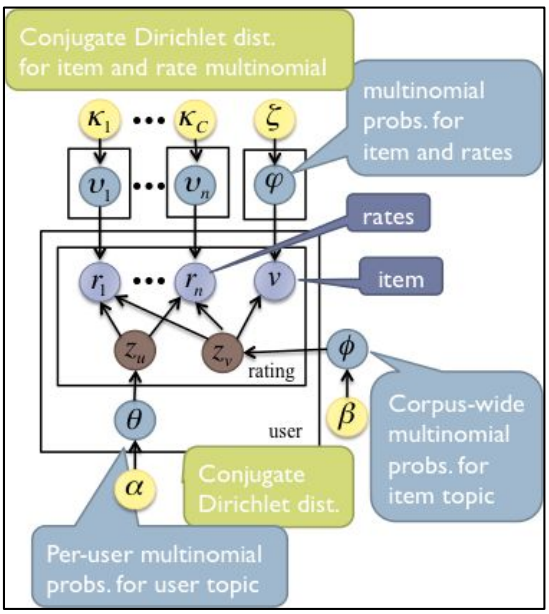
推薦システムに適したトピックモデルの構築と学術情報推薦システムへの応用に取り組んだ。主な研究成果として、①利用者の多様な評価値を活用するモデルの構築、および、②研究者推薦システムの構築があげられる。

①情報推薦に適したトピックモデル

トピックモデルは、通常、文書を語の集合とみなし、潜在トピックを用いて文書を生成する確率モデルとして構築される。このモデルは、生成するデータや潜在構造を変更することで多様なデータや分析タスクに適したモデルに拡張することができる。

情報推薦システムでは、利用者とアイテムという２種類の情報を核としたデータを扱うことになる。そのため、潜在構造として、これらの２つの情報に対応するトピックを導入する必要がある。

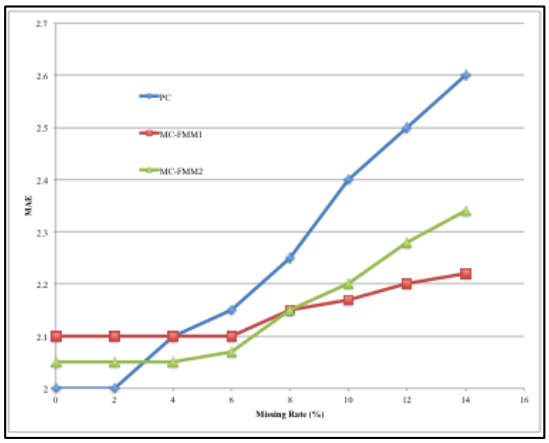
一方、観測されるデータとしては、利用者やアイテムの特性を表す属性および利用者がアイテムに付与する評価値を生成するモデルが必要になる。そこで、本研究では以下に示すトピックモデルを考案した。




図において Zu,Zv はそれぞれ利用者およびアイテムのグループを表す潜在トピックを表している。また、このモデルでは、利用者が多様な観点でアイテムに評価値を与えることを想定しており、r1～rn は各視点での評価値を表している。従って、観測できるデータは、利用者とアイテムの組に対して、n 個の評価値が与えられている。一方、潜在情報としては、利用者とアイテムのトピックがそれぞれ独立に与えられる。モデルのパラメタとして、各利用者のトピック分布θ、アイテムのトピック分布φ、利用者とアイテムの潜在トピックの組に対する各評価値の確率分布 $\upsilon_k$ が用いられる。また、これらの確率分布に対してそれぞれ事前分布が定義されている。

これらのモデルのパラメタを訓練データより推定するために、Gibbs Sampling を用いたアルゴリズムを開発した [2]。

潜在トピックモデルを用いた情報推薦法の特徴のひとつに、よりスパースなデータに適用できる点が上げられる。本研究では、この効果をはかるため、情報推薦システムの評価によく用いられる映画の推薦を題材とした評価用コーパスを用いて考案した手法を評価した。




上記のグラフは、横軸にデータのスパースさを、縦軸に情報推薦の性能を表している。性能値としては、この分野で標準的に用いられる Mean Absolute Error (MAE)を用いており、値が小さいほど推薦の性能がよことを表している。また、ベースラインとして、ピアソンの相関係数を用いた推薦法（青）を用いた。提案手法については、前述したモデルに加え、潜在構造を少し変更したモデル（緑、赤）を用いた。グラフに示されるように、トピックモデルを用いた推薦法は、スパースなデータ（グラフの右側）に対して、ベースラインの手法より良好な推薦性能を得られることがわかる。

②研究者推薦への応用

近年、学術電子図書館においても情報推薦の技術を取り入れる試みがなされている。これまでの研究では、学生や論文執筆を行っている研究者に対する関連論文の推薦という課題に取り組んだものが多かった。本研究では、共同研究を支援・促進するための技術として、研究者の推薦法についての研究を行った[11]。この研究では、まず、研究者間の類似度をはかるための特徴量として、(1)発表文献、(2)共著ネットワーク、(3)引用ネットワーク、(4)研究者の分野への貢献度、という４つの観点から特徴量の検討を行った。そ

して、発表論文から抽出した研究者の研究領域の類似度(ContentSim)、研究者の所属機関レベルの類似度（OrgRS)、被引用数に基づいた研究者の貢献度(I.rate)、貢献度の新鮮さ(ActiveScore)を特徴量とした共同研究者の推薦システムを試作した。

評価実験では、まず、インターネットより1,266,790 件の学術論文を収集し、著者とその所属機関を抽出した。論文に含まれていた研究者数は 807,005 名であった。次に上記の特徴量を計算した。評価にあたっては、共著論文を発表した研究者は共同研究を行ったとみなし、将来、共著論文を発表する研究者をシステムが推薦できた場合に、推薦が成功したと考え、その性能を評価した。

Table I
ACCURACY OF RECOMMENDATIONS WITH DIFFERENT FEATURES

| Features | Precision | Recall | Average Precision |
|---|---|---|---|
| *ContentSim (Baseline)* | **0.5113** | **0.7896** | **0.5328** |
| *ContentSim, OrgRS* | 0.9079 | 0.4367 | 0.8039 |
| *ContentSim, OrgRS, I.Rate* | 0.9079 | 0.4367 | 0.8039 |
| *ContentSim, OrgRS, I.Rate, ActiveScore* | 0.8792 | 0.4953 | 0.8122 |
| *OrgRS* | **0.9133** | **0.4335** | **0.8048** |
| *OrgRS, I.Rate* | **0.9133** | **0.4335** | **0.8042** |
| *OrgRS, I.Rate, ActiveScore* | **0.8864** | **0.4446** | **0.8113** |

なお､性能評価値には､precision と recall を用いた。上記の表は、各特徴と推薦の性能の関係を表している。まず、テーブルの一行目に示される発表論文の類似度(ContentSim)は、共同研究者推薦のための有効な特徴であることがわかる。また、表の 5 行目に示される所属機関の関連性も有効な特徴量であることがわかる。研究機関間での交流・提携がある場合は、共同研究を進め易いことに起因しているものと考えられる。特徴量を組み合わせた場合、所属機関の関係性(OrgRs)と研究者の研究分野への貢献度(I.Rate)を組み合わせることによって高い推薦精度を達成することができた。

以下の表は、推薦する研究者の数と推薦性能の関係を表している。表から読み取れるように、推薦数を変更しても、所属機関の関係性（OrgRs）および研究者の貢献度(I.Rate)が有効な特徴量であることがわかる。

Table II
QUALITY OF THE TOPN RECOMMENDATIONS USING DIFFERENT FEATURES

| Features | Quality of Collaborations | | | | |
|---|---|---|---|---|---|
| | T@10 | T@20 | T@30 | T@40 | T@50 |
| *ContentSim (Baseline)* | **19.64** | **36.89** | **50.50** | **65.68** | **95.30** |
| *ContentSim, OrgRS* | 74.22 | 140.30 | 213.95 | 275.75 | 398.82 |
| *ContentSim, OrgRS, I.Rate,* | 74.23 | 140.31 | 213.95 | 275.75 | 398.80 |
| *ContentSim, OrgRS, I.Rate, ActiveScore* | 102.04 | 175.56 | 233.97 | 292.16 | 446.19 |
| *OrgRS* | **91.57** | **154.52** | **221.18** | **278.60** | **370.69** |
| *OrgRS, I.Rate* | **91.56** | **154.52** | **221.18** | **278.59** | **370.67** |
| *OrgRS, I.Rate, ActiveScore* | **178.75** | **349.76** | **469.04** | **585.74** | **662.89** |

## 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 4 件）


① Manabu Ohta, Ryohei Inoue, Atsuhiro Takasu. Empirical Evaluation of CRF-Based Bibliography Extraction from Research Papers. International Journal on Computer Science and Information Systems, Vol.7, No.2, pp. 18 - 31, 2012. [査読有] http://www.iadisporatl.org/ijcsis/papers/2012150102.pdf

② Tomonari Masada, Atsuhiro Takasu, Yuichiro Shibata, Kiyoshi Oguri. Clustering Documents with Maximal Substrings. Lecture Note in Business Information Processing, LNBIP 102, pp.19-34, 2012.[査読有], DOI: 10.1007/978-3-642-29958-2_2

③ Tomonari Masada, Atsuhiro Takasu, Yuichiro Shibata, Kiyoshi Oguri. Semi-supervised Bibliographic Element Segmentation with Latent Permutations. Lecture Note in Computer Science: Digital Libraries: For Cultural Herritage, Knowledge Dissemination, and Future Creation, LNCS 7008, pp.60-69, 2011. [査読有] DOI:10.1007/978-3-642-24826-9_11.

④ Tomonari Masada, Atsuhiro Takasu, Yuichiro Shibata, Kiyoshi Oguri. Steering Time-Dependent Estimation of Posteriors with Hyperparameter Indexing in Bayesian Topic Models. Lecture Note in Artificial Intelligence: Advances in Knowledge Discovery and Data Mining, LNAI 6634, Part I, pp.435-447, 2011. [査読有], DOI:10.1007/978-3-642-20841-6_36.


〔学会発表〕（計 9 件）


⑤ Tin Huynh, Atsuhiro Takasu, Tomonari Masada, Kiem Hoang. Collaborator Recommendation for Isolated Researchers. in Proc. of MAW14, pp.639 - 644, May 15, 2014. [full paper、査読有], Victoria(Canada)

⑥ Manabu Ohta, Daiki Arauchi, Atsuhiro Takasu, Jun Adachi. Empirical Evaluation of CRF-Based Bibliography Extraction from Reference Strings. In Proc. DAS 2014, pp. 287 – 292, April 10, 2014. [Poster 査読有], Tours(France)

⑦ Padipat Sitkrongwong, Saranya Maneeroj, Atsuhiro Takasu. Latent Probabilistic Model for Context-Aware Recommendations. in Proc. of WI2013, pp.95 - 100, November 18, 2013. [short

paper、査読有], Atlanta(USA)

⑧ Warat Chalempornpong, Saranya Maneeroj, Atsuhiro Takasu. Rating Pattern Formation for Better Recommendation. in Proc. of RSMD2013, pp.146 - 151, August 30, 2013. [full paper、査読有], Prague(Czech)

⑨ Pannawit Samatthiyadikun, Atsuhiro Takasu, Saranya Maneeroj. Bayesian Model for a Multicriteria System with Support Vector Regression. in Proc. of IRI2013, pp.38 - 45, August 15, 2013. [full paper、査読有], San Fransisco(USA)

⑩ Pannawit Samatthiyadikun, Atsuhiro Takasu. Extended Bayesian Model for Multi-criteria Recommender System. 情報基礎とアクセス技術研究会, January 11, 2013. [full paper、査読無], 宮崎JA-AZMホール(宮崎県宮崎市)

⑪ Tomonari Masada, Atsuhiro Takasu. Extraction of Topic Evolutions from References in Scientific Articles and Its GPU Acceleration. in Proc. of CIKM 2012, pp.1522 - 1526, November 11, 2012. [short paper、査読有], Maui(USA)

⑫ Pannawit Samatthiyadikun, Atsuhiro Takasu, Saranya Maneeroj. Multicriteria Collabortive Filtering by Bayesian Based user Profiling. in Proc. of IRI2012, pp.124 - 131, August 8, 2012. [full paper、査読有], San Fransisco(USA)

⑬ Atsuhiro Takasu. A Multicriteria Recommendation Method for Data with Missing Rating Scores. in Proc. of ICDKE 2011, pp.72 - 79, September 7, 2011. [full paper, 査読有], Milan(Italy)

〔その他〕
受賞

① Outstanding Paper Award at International Conference on Information Systems (IS 2012), Empirical Evaluation of CRF-based Bibliography Extraction from Research Papers, Manabu Ohta, Ryohei Inoue and Atsuhiro Takasu, 2012.3.

６．研究組織

(1)研究代表者
高須　淳宏（TAKASU, Atsuhiro）
国立情報学研究所・コンテンツ科学研究系・教授
研究者番号：90216648

(2)研究分担者
正田　備也（MASADA, Tomonari）
長崎大学・工学研究科・准教授
研究者番号：60413928

(3)連携研究者
深川　大路（FUKAGAWA, Daiji）
同志社大学・文化情報学部・助教
研究者番号：10442518