様式 C-19

科学研究費補助金研究成果報告書

平成２３年５月１５日現在

機関番号：１４３０１
研究種目：特定領域研究
研究期間：2006～2010
課題番号：１８０４９０４１
研究課題名（和文）情報爆発に対応するコンテンツ融合と操作環境融合に関する研究

研究課題名（英文）Contents Fusion and Seamless Search for Information Explosion

研究代表者
田中　克己（TANAKA KATSUMI）
京都大学・大学院情報学研究科・教授
研究者番号：００１２７３７５

研究成果の概要（和文）： ウエブからの同位語等の概念知識の抽出，ウエブ検索クエリの意図推定・自動質問修正，ウエブ情報の信憑性分析，ユーザインタラクションやウエブ 1.0 情報とウエブ 2.0 情報の相互補完による検索精度改善に関する技術開発を行った．

研究成果の概要（英文）：We developed methods for (1) extracting ontological knowledge from Web such as coordinate terms, (2) intent detection and automatic modification of Web queries, (3) Web information credibility analysis, (4) improving retrieval effectiveness by user interaction and mutual complementation of Web1.0 and 2.0 information.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2006 年度	16,600,000	0	16,600,000
2007 年度	16,100,000	0	16,100,000
2008 年度	16,100,000	0	16,100,000
2009 年度	17,000,000	0	17,000,000
2010 年度	17,000,000	0	17,000,000
総　計	82,800,000	0	82,800,000

研究分野：情報学
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：情報爆発，情報検索，WWW，ヒューマンインタフェース，情報の信頼性

１．研究開始当初の背景

コンテンツの検索や利活用を行うために，Web1.0 情報の検索を行う Web 検索エンジンや，ユーザ生成コンテンツであるブログなどが台頭していた．学術的，産業的には，Web やマルチメディアのための次世代検索エンジンの開発が望まれていた．また，Wikipedia，ソーシャルブックマーク，twitter，SNS などのソーシャルメディアやユーザ投稿型動画サイトなどは未だ萌芽期の段階にあった．Web 情報の信頼性については意識が低く技術開発などは行われていなかった．さらに，Web，TV コンテンツ，メール，オフィス文書の操作ソフトは個別に提供され，コンテンツ検索操作を効率的に行うための障害となっていた．

２．研究の目的

種々の情報源からの多様なメディア情報を効果的に統合して利活用・管理する技術が不可欠であると考え，本研究では，異種コンテンツの横断的検索・統合や，コンテンツの操作環境と閲覧検索環境の効果的な統合を実現するための基盤的ソフトウェア技術を開発し，情報爆発に対処可能な新しいコンテンツ検索・融合技術を得ることを目的とした．

３．研究の方法

コンテンツの統合検索，高水準のユーザ・コンテクストに着目した情報検索，コンテンツ統合とシームレスな検索という着眼点に基づき，理論的研究，システム的な研究開発を行うというアプローチをとった．

４．研究成果

(1) Web からの知識獲得とサーチ

検索キーワードに対する典型的な話題語・同位語・上位語・下位語といった知識を Web 全体から近似的に高速抽出する新しい技術を開発した．これは両方向言語構文パターンを用いた知識獲得手法で，知識の発見には，既存の検索エンジンがもつ索引データ（ページのタイトル，スニペット，URL）のみを利用するものである．与えられた語に対して特定の関係にある語（上位語や下位語等）が対象語と言語構造的にどのように共起するかということに注目し，2 種類の異なる方向性を持つ言語パターンを用意し，それらに適合するかどうかということを判断することで知識の抽出が行われる．

たとえば，語<query>の同位語を発見するには，前構文パターン「<target>や<query>」と，後構文パターン「<query>や<target>」を用いる．<target>が発見したい語が現れる部分であり，<target>がパターンの最初に現れるものを前構文パターン，最後に現れるものを後構文パターンと呼ぶ．<target>を除く文字列をクエリとして，検索エンジンから検索結果を取得することで，<target>の部分に該当する語を発見することができる．

語の発見において，特に日本語や中国語においては，語の切れ目を判別することが重要であるが，前構文パターンによって<target>の終端部が，後構文パターンによって<target>の開始部が特定されるため，これらの両方向のパターンの両者に適合する語を抽出することで，語の切り出しが容易に精度良く行われる．この手法は発見したい語の種類や言語を問わず利用可能である．オンデマンドに取得した知識をバネモデルを用いて視覚的に表示するシステムの作成も行った（図 1）．

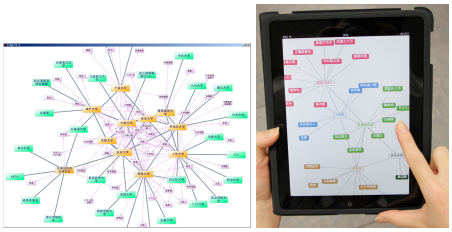

図 1：Web からの同位語／話題語知識の獲得

また，ブログ検索やニュースアーカイブ検索のような，検索対象のコンテンツの生成日時をクエリで指定できる検索エンジンを利用することで，語の関係の変化を抽出することができるシステムを作成した．

(2) Web2.0 による Web1.0 検索の精度改善

Web2.0 サイトにおけるタグ情報をマイニングすることで既存の Web 検索エンジンの適合率・再現率を向上させる手法を提案した．

Web 情報検索におけるランキングを最適化する目的で，ウェブページの適合性を判断するうえでソーシャルブックマークのデータに着目し，このデータをマイニングすることで既存のページ評価尺度と比較した上で，これまでの検索エンジンでは検索が難しかったページの検索を可能にする手法を提案した．具体的には，ソーシャルブックマークにおけるページのブックマーク数や，ブックマークの際に使用されたタグといった情報をページへの一種のアノテーションと見なすことで，これを用いた検索結果の対話的に再ランキングを可能にし，従来は検索結果からの発見が難しかったページの発見を容易にした．また，ソーシャルブックマークにおいて特定の時期にブックマーク量が増加するページを，その時期に特に有用な情報が載っていると仮定することで，検索者が検索を行う時期に適応した検索結果のランキングを可能にした．さらに，ページをブックマークしているユーザ間の参照構造を分析し，集団としての特性を明らかにすることで，ページの社会的なインパクトをより正確に反映するモデルと指標を提案した．

現在の Web 画像検索エンジンは「金閣寺」や「富士山」といった具体物を表す語で検索した場合の検索精度は高いが，「夏」や「平和」のような抽象的な概念を表す語で検索した場合にはその精度（再現率）が著しく低下する．それに対して，たとえば抽象的な語である「春」を検索語にするとき，この語を「クロッカス」，「チューリップ，庭」，「桜」などといった具体的な語集合に置き換えてから既存の画像検索エンジンを用いて検索することで，検索精度を向上させることができる（図 2 参照）．ここで重要なのは検索語を抽象的な語から具体的な語の集合に変換する過程である．そこで抽象的な概念を表す語をタグとして付与された画像のソーシャルタグ集合に置き換えるとき，選択する置き換え語が本来の検索語である抽象的な語をどのくらい連想させるかをあらかじめ用意した相関ルールを用いて決めることで検索精度を向上させた．

図 2：Web2.0 データによる抽象語クエリから具体語集合への変換とそれによる画像検索

(3) 関係の類似性に基づく検索とアナロジー検索

関係が与えられた時に，類似関係を Web から発見する方法，および，関係の類似性によるアナロジー検索を開発した．類似関係検索

は，入力として語 A，B，Cが与えられた時に，A と B で成り立つ関係と C と D で成り立つ関係が類似するような語 D を Web から発見する．

我々は Web ページの索引情報を情報源とし，語の共起と言語パターンを用いて類似関係検索を実現する手法について開発した．

また，アナロジー検索では「京都におけるある飲食店 A は東京においてどの飲食店に当たるか」という検索を行うことができる．飲食店間の類似度を考えることで例示検索は実現できるが，我々は更に，選ばれた例と選ばれなかった例の関係，例が選択された地域（京都）と検索対象となる地域（東京）の関係を考慮することで，より精度よく，また異なる地域間での例示検索を可能にした（図 3）．

図 3：アナロジー検索

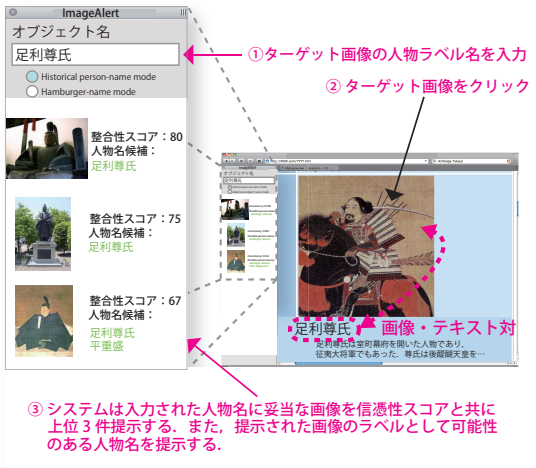

図 4：画像信憑性分析システム ImageAlert

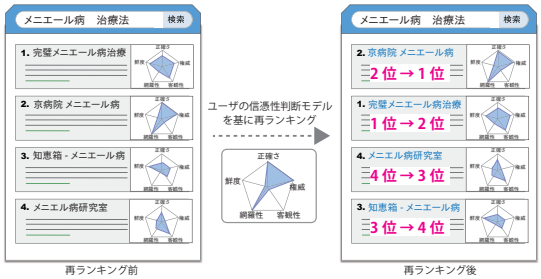

図 5：ユーザ信憑性判断指標による再ランク

(4) ウェブ情報の信憑性分析に関する研究

信憑性指向の情報検索・閲覧を実現するために，ウェブ情報の信憑性分析モデル，情報検索時におけるユーザ支援方法，ウェブからの信憑性判断材料の検索・集約に関する技術開発を行った．

ウェブ情報の種類や評価観点によって異なる信憑性を集合的アプローチによって評価するためのモデルを提案した．提案モデルでは，任意の観点からウェブ情報の信憑性を評価するために，評価対象であるウェブ情報をデータ対として表現し，関連するデータ対間のサポート関係の強さを分析する．このとき，「Supportive なデータ対を多く持つデータ対は信憑性が高い」という仮説のもと，データ対として表現されたウェブ情報の信憑性評価を行う．信憑性の評価観点に応じてデータ対の構成，データ対間のサポート関係の定義を行うことで様々なウェブ情報の信憑性を評価することが可能となる．提案モデルを画像信憑性の評価に適用し，ウェブページを閲覧中に気になる画像の信憑性を検証できるアプリケーション ImageAlert を開発した（図 4）．

さらに，ユーザの検索結果に対する信憑性フィードバック情報をもとにユーザの信憑性判断モデルを推定し，それを用いてウェブ検索結果を再ランキングするシステムを開発した（図 5）．提案システムでは，ウェブ検索結果を信憑性の主要な判断指標ごとにスコアリングする．ユーザは可視化された信憑性判断指標ごとのスコアを確認し，システムに各検索結果の信憑性をフィードバックすることができる．提案システムによって，ユーザは自身の信憑性判断指標に応じて，信憑性の高いページを大量のウェブページの中から効率よく検索することが可能となる．

近年のユーザ投稿型サイトやオンライン広告サービスの普及により，Web 上には一般のユーザや業者によって投稿されたコンテンツが多数存在している．これらのコンテンツで扱われる料理や旅行ツアーといったオブジェクトは，書き手によって恣意的に名前が付けられるため，同一オブジェクトに対して無数の名前が存在する．そのため，ある Web ページがどのオブジェクトについての記述であるかを識別することは容易ではない．また，これらのコンテンツでは，記述対象のオブジェクトをより魅力的に見せるために，様々な修飾表現が用いられる．例えば，“本格カレー” という名前の料理レシピや，“大感動！上海 4 日間” といった名前の旅行ツアーが存在する．しかし，修飾表現を含むオブジェクトであっても，実際にはその修飾表現と適合していないオブジェクト（誇張）や，逆に，ある修飾表現を含まないオブジェクトであっても，その修飾表現と適合しているオブジェクト（隠れ適合）も多数存在する．こ

れらは，クエリ内のキーワードを含むオブジェクトを検索結果に表示する既存の検索エンジンでは，適合率と再現率の低下の原因となる．

本研究ではオブジェクトに対する修飾表現の適合性という，人々がそのオブジェクトに関する記述を閲覧した際に，その修飾表現が相応しいと考える度合いを示す指標を導入した．オブジェクトに対する修飾表現の適合性分析を行うために，修飾表現と適合する語と相反する語を抽出する手法を開発した．オブジェクトに対する修飾表現の適合性という指標は，修飾表現を含むクエリでの検索精度の向上や，誇張されたオブジェクトの発見などに利用できる．料理レシピと旅行ツアーを対象として実験を行い，検索タスク・誇張発見タスクにおける精度評価を行った．その結果，ほぼ全ての手法において，既存の検索エンジンに基づくベースライン手法などよりも，提案手法の方が優れていることを示した．

(4)サーチインタラクションの研究

ユーザ主導のサーチインタラクションにもとづく情報検索技術の研究として，語ベースフィードバックに基づく検索結果の再ランキング，Q&A コンテンツからの主観的ファセットの抽出に関する研究を行った．

検索エンジンの膨大な検索結果を，検索結果リストや俯瞰的に重要語を表示したもの（タグクラウド）の中から任意の語を強調・削除することで迅速に再ランキングできる方式 Rerank（http://rerank.jp）を開発した（図 6）．またシステムを Yahoo！Japan と共同でシステムの開発を行い，Yahoo!ラボで公開し実証実験を行った．提案システムを利用することで，ウェブ検索エンジンのランキングに左右されることなく効率的に情報を発見することが可能となる．

図 6：語ベースフィードバックに基づく再ランキングシステム Rerank.jp

さらに，ユーザ側の簡単なインタラクションで任意の検索サービスの検索結果ページの構造抽出を可能にする技術を実現し，さまざまな検索サービスで再ランキングが使用できるシステム RerankEverything を開発した．提案システムはユーザが検索結果ページに対してインタラクションすることで，動的に検索結果の属性値を抽出し再ランキングに利用する．これにより，検索サービス側の検索機能に依らず，ユーザ主導の観点で検索結果を閲覧することが可能となる（図 7）．

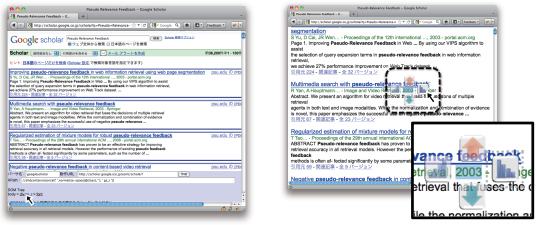

図 7：RerankEverythingを用いて論文検索サービスの検索結果の構造を抽出し，論文情報を再ランキングする様子

次に，コミュニティ型の Q&A コンテンツの質問・回答情報を利用し，主観的ファセットを抽出することで，ウェブ検索ユーザのクエリ修正を支援する手法を実現した．提案システムでは Q&A コンテンツに出現する「有名な寺社」や「美味しい和菓子」といった主観的ファセット情報に焦点を当て，質問・回答情報からそうしたファセットを抽出し，ウェブ検索ユーザに対して推薦する．

システムはまずユーザからクエリを受け取ると関連する Q&A コンテンツを検索し，あらかじめ用意した言語パターンを使用して質問中に出現する主観的ファセットを抽出する．その後，得られた主観的ファセットと，回答に出現する「金閣寺」や「銀閣寺」といったエンティティ名の共起関係を評価することで，検索クエリと関連するファセットを動的に発見する． 提案システムを利用することで，「京都観光」，「インド料理」といった曖昧な検索クエリに対しても効果的にクエリを推薦することが可能となる（図 8）．

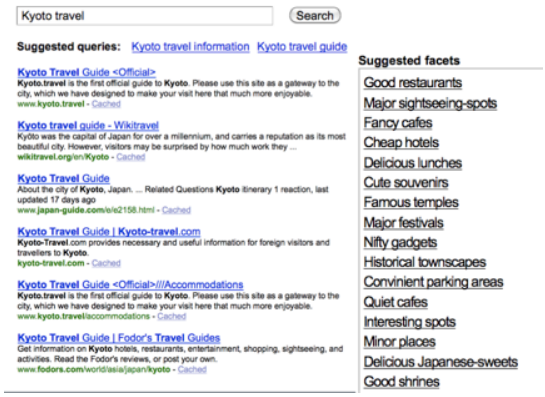

図 8：Q&A コンテンツから抽出された主観的ファセットの例

(5)オフィス文書の意味的検索技術の研究

Google デスクトップサーチ等の既存の全文検索システムの大きな欠点として，検索条件を定義する際にユーザが検索語の意味を定義できないために検索の適合率や再現率が低いことがあげられる．本研究ではこの欠点に対応した意味的検索手法の研究を行い，本手法を用いた MS Word と Excel 等のオフィス文書の意味的検索システムを開発した．

文書管理者は各種の業務文書の中の検索用語句（例えば，顧客名・発売日・製品名・単価・合計金額等）の意味を表す意味タグを意味タグスキーマとして定義し，各種の代表文書の中の検索用語句に意味タグスキーマの要素を対応付けておく．検索用語句の抽出プログラムはその代表文書と構造的・意味的に類似した文書の中から自動的に検索用語句と意味タグを XML データとして出力する．検索インデックス作成プログラムはこれらの XML データを読んで，関係データベースに保存した検索インデックスの更新を行う．

これにより，ユーザは意味タグを意識せずに検索条件を定義できる．開発したシステムは，検索対象となる文書の種類の意味タグを検索インデックスの意味タグ定義表から読み出し，HTML の検索フォームを生成する．文書検索プログラムは検索フォームで定義されたユーザの検索条件から検索クエリを自動的に生成し，検索条件に満たす文書の一覧を出力する．文書一覧には文書の検索用語句，意味タグ，文書の保存先，文書へのリンクがあるので，ユーザはこの情報を見て当該文書を確認することが可能になる．検索速度測定結果は，10 万件の文書から 10 件のターゲット文書の検出時間は約 0.03 秒であった．

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 52 件）

[1] Hiroaki Ohshima , Katsumi Tanaka : High-speed Detection of Ontological Knowledge and Bi -directional Lexico-Syntactic Patterns from the Web, Journal of Software , 5(2) 195-205 (2010)(査読有り)

[2] Katsumi Tanaka et al.: Improving Search and Information Credibility Analysis from Interact ion between Web1.0 and Web2.0 Content. J. Software 5(2): 154-159 (2010) (査読有り)

[3] 山本 祐輔,田中 克己: データ対間のサポート関係分析に基づく Web 情報の信憑性評価，情報処理学会論文誌 データベース，3(2)(TOD46),61-79 (2010) (査読有り)

[4] 高橋 良平,小山 聡,大島 裕明,田中 克己: Web テキストと修飾表現との適合度判定手法，日本データベース学会論文誌(DBSJ Journal)，9(1),41-46 (2010) (査読有り)

[5] 山本 岳洋,中村 聡史,田中 克己: RerankEverything:ランキング結果閲覧のための柔軟な再ランキングインタフェース，情報処理学会論文誌:データベース，3(4)(TOD48)，48-64 (2010) (査読有り)

[6] 加藤 誠,大島 裕明,小山 聡,田中 克己: 関係の類似性に基づく Web からのオブジェクト名検索，情報処理学会論文誌 データベース (TOD 42)，2(2)，110-125 (2009) (査読有り)

[7] 山家 雄介,中村 聡史,アダム ヤトフト,田中 克己: ソーシャルブックマークの特性分析とそれに基づく Web 検索結果の再ランキング手法，情報処理学会論文誌データベース，1(1)，88-100 (2008) (査読有り)

[8] 山本 岳洋,中村 聡史,田中 克己: Rerank-By-Example:編集操作の意図伝播によるウェブ検索結果のリランキング，情報処理学会論文誌:データベース，49(TOD37) (2008) (査読有り)

[9] 大島 裕明,小山 聡,田中 克己: Web 検索エンジンのインデックスを用いた同位語とそのコンテキストの発見,情報処理学会論文誌:データベース，47(SIG(TOD32)) (2006) (査読有り)

〔学会発表〕（計 267 件）

（国際会議論文 115 件,国内学会論文 131 件，招待講演・基調講演 21 件）

[1] Yusuke Yamamoto, Katsumi Tanaka : Enhancing Credibility Judgment of Web Search Results , ACM CHI2011, 1235-1244, May 2011

[2] Yusuke Yamamoto, Katsumi Tanaka : ImageAlert: Credibility Analysis of Text-Image Pairs on the Web , ACM Symposium On Applied Computing (SAC 2011), 1724-1731, March 2011.

[3] Makoto P. Kato, Hiroaki Ohshima , Satoshi Oyama, Katsumi Tanaka: Search as If You Were in Your Home Town: Geographic Search by Regional Context and Dynamic Feature-Space Selection, 19th ACM CIKM 2010,1541-1544 (2010)

[4] Takehiro Yamamoto, Satoshi Nakamura, Katsumi Tanaka: RerankEverything: a Reranking Interface for Browsing Search Results, WWW2010, 1209-1210

(2010)
[5] 山本 岳洋,中村 聡史,田中 克己: QA コンテンツからの観点抽出とそれにもとづくウェブ検索結果の再ランキング, Web とデータベースに関するフォーラム(WebDB Forum2010), 2A-2 (2010)
[6] Takehiro Yamamoto, Satoshi Nakamura, Katsumi Tanaka: Reranking and Classifying Search Results Exhaustively Based on Edit-and-Propagate Operations , DEXA2009, LNCS, 855-862 (2009)
[7] Hiroaki Ohshima, Adam Jatowt, Satoshi Oyama, Katsumi Tanaka: Seeing Past Rivals: Visualizing Evolution of Coordinate Terms over Time , 10th WISE2009, LNCS, 167-180 (2009)
[8] Makoto P. Kato, Hiroaki Ohshima , Satoshi Oyama, Katsumi Tanaka: Query by analogical example: relational search using web search engine indices. ACM CIKM 2009: 27-36
[9] Makoto Kato, Hiroaki Ohshima, Satoshi Oyama, Katsumi Tanaka : Can Social Tagging Improve Web Image Search?, Proc.WISE2008, LNCS, 5175, 235-249 (2008) (Best Paper Award).
[10] Somchai Chatvichienchai , Katsumi Tanaka: Bring Precision and Access Control to Business Document Search, Proc. 9t h ACIS International Conference (SNPD2008), 557-562 (2008)
[11] Yusuke Yanbe, Adam Jatowt, Satoshi Nakamura, Katsumi Tanaka: Can Social Bookmarking Enhance Search in the Web?, Proc. of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL2007),107-116 (2007) (Nominated for Vannevar Bush Best Paper Award)
[12] Somchai Chatvichienchai, Jinsan Lin, Katsumi Tanaka : Towards XML -Based Index for Effective Search of Office Documents, Proc. of the Int. Conf. on Business and Information (BAI2007) (2007)

〔図書〕(計 12 件)

[1] Yusuke Yanbe, Adam Jatowt, Satoshi Nakamura, Katsumi Tanaka : Social Bookmarking and Web Search , San Murugesan, Handbook of Research on Web 2.0, 3.0, and X.0: Technologies, Business, and Social Applications, IGI Globals, 242-259 (2009)
[2] Katsumi Tanaka, Takashi Matsuyama, Ee-Peng Lim,Adam Jatowt: Proceedings of the 2nd ACM Workshop on Information Credibility on the Web, WICOW 2008, ACM (2008)
[3] Akiyo Nadamoto, Yoshiharu Ishikawa, Masaru Kitsuregawa, Katsumi Tanaka: Proceedings of the internation al Workshop on Information -Explosion and Next Generation Search, INGS2008, Shenyang, China, April 26 -27, IEEE Computer Society (2008)

〔その他〕
ホームページ:
http://www.dl.kuis.kyoto-u.ac.jp

６．研究組織

(1)研究代表者
田中 克己 (TANAKA KATSUMI)
京都大学・大学院情報学研究科・教授
研究者番号：00127375

(2)研究分担者
チャットウィチェンチャイ ソムチャイ (CHATVICHIENCHAI SOMCHAI)
長崎県立大学・国際情報学部・教授
研究者番号：00382440

田島 敬史 (TAJIMA KEISHI)
京都大学・大学院情報学研究科・准教授
研究者番号：60283876

小山 聡 (OYAMA SATOSHI)
北海道大学・大学院情報科学研究科・准教授
研究者番号：30346100

中村 聡史 (NAKAMURA SATOSHI)
京都大学・大学院情報学研究科・特定准教授
研究者番号：50415858

手塚 太郎 (TEZUKA TARO)
筑波大学・大学院図書館情報メディア研究科・准教授
研究者番号：40423016

ヤトフト アダム (JATOWT ADAM)
京都大学・大学院情報学研究科・特定准教授
研究者番号：00415861

(3)連携研究者

大島 裕明 (OHSHIMA HIROAKI)
京都大学・大学院情報学研究科・助教
研究者番号：90452317