様式 C-19

科学研究費補助金研究成果報告書

平成 23 年 4 月 30 日現在

機関番号：62615
研究種目：特定領域研究
研究期間：2006～2010
課題番号：18049069
研究課題名（和文）　情報爆発時代の情報検索基盤技術
研究課題名（英文）　Infrastructure for Information Retrieval in Information Explosion Era
研究代表者
安達 淳（Adachi Jun）
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：80143551

研究成果の概要（和文）：　無秩序に拡大する大量情報の中から必要な情報を的確に取り出し、わかりやすく提示するための情報リンケージ基盤技術の研究を進めた。類似する情報を効率的に見つける検索索引、半構造データのための木類似度計算、データベースを横断して適用可能なレコード同一性判定などの研究課題に取り組み、データ構造やアルゴリズムを提案して有効性を示した。また、論文や研究者を対象とする大規模なリンケージシステムを開発して、NII の学術コンテンツ基盤上で実証した。

研究成果の概要（英文）：　This research focused on fundamental techniques for information linkage to enable users to quickly identify and learn necessary information in a chaotic and fragmented mass of information. The proposed techniques include search indexes for efficient similarity search, an approximate calculation of tree similarity for large scale semi-structured data, a versatile record matching method for heterogeneous database linkage. We also developed a linkage system for scientific papers and researchers the usefulness of which we demonstrated using a nationwide academic content service provided by NII.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2006 年度	16,700,000	0	16,700,000
2007 年度	16,800,000	0	16,800,000
2008 年度	16,700,000	0	16,700,000
2009 年度	16,100,000	0	16,100,000
2010 年度	16,000,000	0	16,000,000
総　計	82,300,000	0	82,300,000

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、情報リンケージ、大規模コーパス、テキスト処理、機械学習

１．研究開始当初の背景

Web 上には日々大量の情報が蓄積され、その量は個々の人間の処理能力を超えて増加していた。その結果、情報取得に要する時間が過剰に増大し、必要な情報はどこかに存在するがその情報に辿り着くのが困難な状況にあった。そこで、このような爆発的に増加する情報に対する効果的な情報検索手法を開発することが重要な課題となっていた。

２．研究の目的

情報は爆発的に増加するように見えるが、その中には重複した情報がかなり含まれており、また、関連する情報が散在していることが、情報の取得を困難にしていると考えられる。そこで、本研究は、インターネット上

で公開される各種テキストや個人・組織が管理する文書を対象として、関連する情報を結び付ける「情報リンケージ」プラットフォームの実現を目的としている。本研究で提案する情報リンケージ技術とは、無秩序に拡大する大量情報の中からその場で自分の必要とする情報を的確に取り出し、わかりやすく提示するための基盤技術を意味し、従来からの情報検索、フィルタリング、質問応答の技術と情報統合を組み合わせ、新しい情報獲得手法として体系化を図ろうとするものである。情報リンケージの対象としては、本研究では特に、本や研究者といった学術情報に含まれる具体物を対象として、(i)Web 上に重複して現れる具体物のリンケージ手法の考案、(ii)リンクづけされた具体物の効果的かつ効率的な検索法、(iii)大規模情報を用いたリンケージシステムの構築と評価を行う。

３．研究の方法

情報リンケージ基盤として、(1)情報収集、(2)特徴選択と抽出、(3)マッチング,(4)情報検索の４との機能の層およびこれらの機能層を横断する形で大規模データを現実的な時間で処理するためのデータ構造とアルゴリズム群からなるアーキテクチャを考え、各層の機能実現およびデータ構造とアルゴリズムの開発を研究代表者、分担者、連携研究者、研究協力者で手分けして研究を進めた。

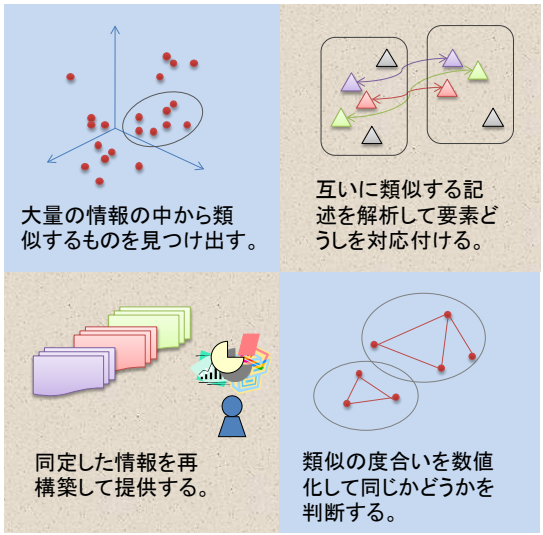

図１　本研究の研究概要

４．研究成果

主な研究成果は、以下の図に示すように、(i)大量の情報の中から類似情報を見つけるデータ構造とアルゴリズムの考案、(ii)複数の要素から構成される学術情報のオブジェクトの間をオブジェクトレベルおよび要素レベルでリンクづける照合法の開発、(iii)類似度に基づいてオブジェクトが同一の実体を表しているかどうかを判定する同定法の検討、および、(iv)これらの技術を用いた大規模リンケージシステムの開発である。

（１）類似検索手法の開発

大量の情報の中から類似するデータを高速に見つけ出すことを目的とした、類似検索索引と近傍ペア探索アルゴリズムの開発に取り組んだ。情報リンケージでは多様な形式のデータを扱うため、本研究は距離空間という極めて基礎的な距離に関する空間を扱った。

類似検索索引

類似検索索引は、事前に索引付けしたデータセットの中から任意のオブジェクトと類似したものを探すのに使われる。この索引は、情報リンケージでは信頼性の高いデータベースと表記揺れや欠落を含むオブジェクトとを結びつける処理に利用する。

類似検索索引は、クエリから距離の遠いオブジェクトを枝刈りし、距離計算やディスクアクセスといった検索コストを削減する。ほぼすべての類似検索索引は、Pivot と呼ばれる参照オブジェクトを使っている。類似検索索引では、Pivot からの距離で空間を部分空間へ再帰的に分割し、木構造の索引を構築する。この索引は、検索処理中では、三角不等式を使って、分割された部分空間を枝刈りするのに利用される。つまり、Pivot の選択手法は、検索索引の構造と枝刈りの性能を決定すると言える。従来の Pivot 選択手法は経験則にもとづいたものや簡単なデータマイニングを組み合わせたものが多かったが、いずれの手法も性能はオブジェクトの分布に依存し、どの分布に対しても検索コストを削減できるわけではなかった。

これに対して、本研究はオブジェクトの分布にもとづいた Pivot 選択手法の方策を提案した。そしてこの方策にもとづく、2 つの新しい類似検索索引、Maximal Metric Margin Partitioning (MMMP)と Pivot Capacity Tree (PCTree)を開発した。

MMMP はまず、データの分布傾向のうち特にクラスタの境界を抽出する。そして、クラスタ形状に基づいて Pivot とその分割距離を決める。MMMP の分割面は、隣り合うクラスタの橋からの距離を最大にするように置かれる。MMMP は偏った分布のデータに対して効果的な手法である。人工の 2 から 30 次元のベクトルと 3 つの実データに対して、iDistance、D-index、および List of clusters の 3 つの先行研究との間で、検索応答時間、距離計算回数、ページアクセス回数の比較を行い、いずれにおいても提案する MMMP が良好な特性を持つことを示した。しかし、課題としてクラスタ化していないオブジェクト空間において効果が小さいことも判明した。

一方、PCTree はデータの分布だけでなく、

索引木のバランスも考慮した手法である。PCTree では、Pivot によって分割される部分空間のバランスと、Pivot による検索時の枝刈りの効果の、2 つを考慮して Pivot を選択する。その結果、PCTree はデータの分布に合わせて索引構造を効果的に変化させている。PCTree は MMMP の索引木が不均衡になりうる欠点を改善した手法だと言える。図 2 に PCTree で Pivot を選ぶ際に検討する領域についての概要図を示す。人工の 2 から 64 次元のベクトルデータと 5 つの実データに対して、GHT、MVP、List of clusters、および SAT の 4 つの先行研究との間で、近傍検索に必要な距離計算回数、そして索引木の高さなどを比較した。その結果、提案手法は様々な分布のデータに対して全般的に有効な索引であることが明らかになった。

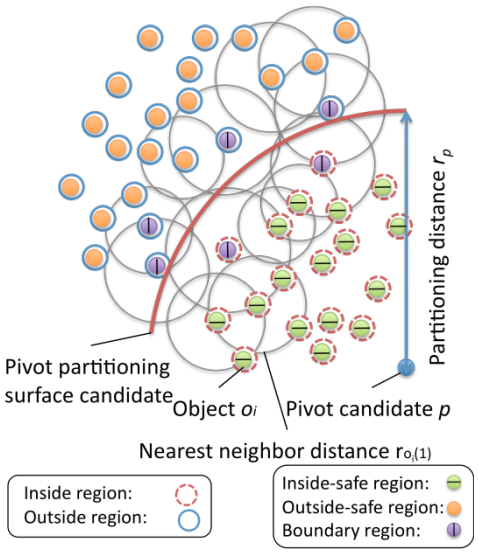

図 2　PCTree の概念図

近傍ペア探索アルゴリズム

近傍ペア探索アルゴリズムは、データセットの中から類似した上位 k 個のオブジェクトのペアを探すのに使われる。このアルゴリズムは、情報リンケージでは表記揺れや欠落を含むデータにおいて類似したものを結びつけて大まかに整理する処理に利用する。

近傍ペア探索アルゴリズムでは、k 番目の類似ペア間の距離の上限値を、初期値∞から類似ペアを見つけるたびに減少させる。さらに、類似検索索引と同様に Pivot からの距離と三角不等式を使って、上限値よりも距離が遠いと判断できるオブジェクトのペアを枝刈りする。この枝刈りは、上限値が小さく、Pivot からオブジェクトまでの距離が分散しているときほど効果が大きい。しかしながら、従来手法は、k 番目の類似ペア間の距離の上限値が収束していく特徴を枝刈り手法に利用していなかった。

これに対して、本研究は適応型空間多分割による分割統治法の k 最近傍ペア探索手法、Adaptive Multi Partitioning (AMP)を提案した。AMP は Pivot からオブジェクトまでの距離が分散している空間から順に分割・統治のステップで k 最近傍ペアを探索する。距離に対するオブジェクトの分散は、距離の分布の歪度をもとに判断する。本手法は、距離に対するオブジェクトの分布が密な空間のほうが、収束した上限値による枝刈りの効果が大きいことを利用している。図 3 に AMP の概要図を示す。3 つの実データに対して、Quickjoin および AMP の分割順序を逆にした手法との間で、距離計算回数の比較を行い、いずれにおいても提案する AMP が良好な特性を持つことを示した。

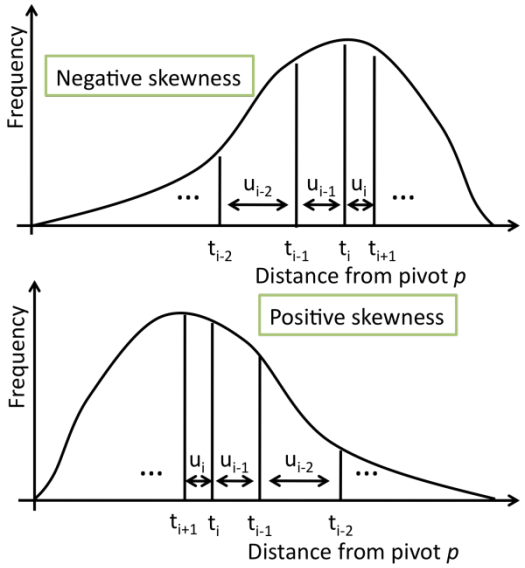

図 3 AMP の概要図

（2）半構造データマッチング

半構造データのマッチングのための基本アルゴリズムとして、順序木のマッチングおよび無順序木のマッチングに取り組んだ。

順序木の近似マッチングアルゴリズム

本研究では，木構造の編集距離を計算する代わりに，木構造から一意に決まる Euler 文字列と呼ばれる文字列について編集距離を計算し，もとの木構造の編集距離を近似するアルゴリズムについて，性能の解析を行なった．その結果，Euler 文字列に少し修正を加える事により，既存のものよりも良い近似度で木構造の編集距離を計算する$O(n^2)$時間アルゴリズムを得た．これは厳密な編集距離を計算する$O(n^3)$時間アルゴリズムよりも速く，文字列の編集距離計算と同等の計算時間コストを持つという点で高速なアルゴリズムといえる．近似度の証明を含む詳細については発表文献[ISAAC06]を参照されたい。

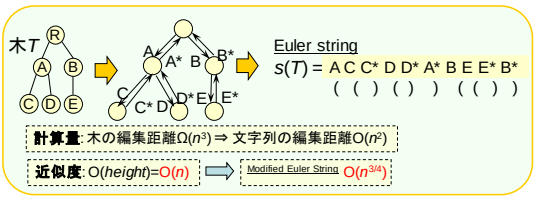

図 4　文字列編集距離を用いた木編集距離

順序木の確率的類似度モデル

木の編集距離に基づいて木構造データのマッチングを行う場合に、挿入や削除等の編集操作のコストを問題に応じて設定することが必要になる。本研究では、このコストを訓練データから推定する確率的なモデルとそのモデルに基づいた編集操作コストの推定法についての研究を行ってきた。この研究では、類似木のペアを一定の確率分布に基づいて生成する generative model を定義し、そのモデルのパラメタを訓練データとして与えられる類似木対から推定することで、編集操作のコストを確率的に定義するものである。木のラベル数が多くなりにつれて、モデルのパラメタが多くなり、過学習に陥りやすくなるため、パラメタのベイズ推定法についての研究を行った。隠れマルコフモデルを用いた木の類似度モデルのパラメタは基本的に多項分布となるため、本年度は、事前分布として Dirichlet 分布を用い、変分ベイズ法を用いてパラメタを推定した。次の図は人工データを用いたモデルの評価結果を表している。本実験では、類似木のペアをランダムに生成し、その一部を訓練データ、残りをテストデータとして用いた。テストデータに含まれる木の集合を木構造データベースと見なし、問い合わせとしてテストデータに含まれる木 T を与え、訓練データから得られた類似度モデル用いて、DB 中の木と T との類似度を計算しランキングする。そのときに、T のペアとなる木が、検索結果の何位以内に入る割合をプロットしたものである。この図には、最尤推定によって得られたモデル(em)とベイズ推定によって得られたモデル(vbem)の検索精度が示されている。上記の人工データを用いた実験では、当初の想定どおりベイズ推定をしたモデルの検索性能が優れていることが示された。

無順序木の近似マッチングアルゴリズム

木の類似度として最も一般的に用いられているのは編集距離である。文字列は木の特殊な場合(葉の数が 1) と考えられるため、木の編集距離は文字列の編集距離の一般化であり、より難しい問題である。もう少し一般化した順序木の場合においても、現在知られている中で最速のアルゴリズムでさえも文字列の場合よりも多くの計算量を必要とする。さらに、無順序木の場合には木マッチングは Max SNP 困難であることが証明されており、さらに難しい問題となる。本研究では，無順序木の編集距離計算のための 2*h*+2 近似アルゴリズムを考案した。ここで、*h* は木の高さの上限を表している。このアルゴリズムでは、まず、比較する２つの木の部分木をすべて列挙し、そのL1 距離をはかる。するとL1 距離と元の無順序木の編集距離との間に以下の関係が成り立つことを示した。

$$\frac{1}{2h+2}\left\|\phi(T_1)-\phi(T_2)\right\|_1 \le TED(T_1,T_2) \le \left\|\phi(T_1)-\phi(T_2)\right\|_1$$

（３）大規模リンケージシステムの開発

情報リンケージの技術を使うことにより、互いに関連する情報どうしを関連付け、再構築して活用することが可能になる。特に、長い年月に渡り人手をかけて構築された大規模なデータベースは信頼性の高い情報源であり、リンケージによる情報集約の有効性が期待される。これに基づき本研究テーマでは、大規模なデータベースを知識として活用する「エンティティ」同定の枠組を提案し、これを実現するための実用的な「リンケージサーバ」の構築手法を検討して、実装・評価を行った。

ここで、本研究における「エンティティ」とは、同定対象であるテキストが指示する実世界の実体」を指している。これは、同定対象がすでにデータベースで管理されている場合には、データベース上でユニークな識別子（ID）を付与されたレコードに対応する。一方、同定対象が明示的にデータベースで管理されていない場合については、エンティティ集合をどのように定めるかは自明ではない。そこで、まず同一と考えられる情報をグループ化した上で、新たに統一的でユニークな識別子を付与する仕組みが必要である。

本研究で実証のターゲットとした学術コンテンツにおいては、前者が「論文」、後者が「論文の著者（＝研究者）」に対応する。以下では、論文と著者それぞれに対するリンケージサーバの構築について報告する。

書誌リンケージサーバの構築

本提案による書誌リンケージサーバでは、長い単位での単語Ｎグラム一致を優先的に検索することにより、数千万件規模のデータベースに対する高速で性能のよい同定を実現している。具体的には、(1)言語解析による前処理、(2)高速なリンケージエンジンによる同定候補抽出、(3)機械学習の適用による同一性の判定、の３つの手順で処理を行う。

従来からデータベース分野においては、データベースの整合性チェックや異種データベース統合の目的で、重複レコードの検出に

関する研究が行われてきた。提案手法はこれをテキストとデータベースの同定に拡張した点が特徴的であり、フォーマットが指定されないテキスト記述をクエリとしてデータベース中の一致レコードを検索することが可能である。テキスト中に複数のレコードへの参照が含まれる場合にも、候補となるレコードそれぞれについて同定処理が行える。

本研究では、提案する論文リンケージシステムを実装し、国立情報学研究所が全国規模で事業サービスを行う論文データベースに対して実証的に適用した。これにより、約 1 千万件の引用文献や成果リストを、総計数千万件の論文データベース上のレコードと対応づけて、著者リンケージや科学計量のための分析用データとして提供するなど、有効性を確認した。また、任意の入力に対して同定をオンラインで行うデモシステムを公開した（http://ci.nii.ac.jp/）。

さらに、提案手法は、OCR による自動文字認識結果など、ノイズを含む入力に対しても頑強である。そこで、OCR や PDF の解析ツールによって切り出されたノイズを含む引用文字列を同定するための論文リンケージ用 API や判定用の GUI を実装し、デモシステム上であわせて提供している。

図 5 OCR 文字列に対するリンケージの例

著者リンケージサーバの構築

本提案による著者リンケージサーバでは、多数の表記揺れやデータの欠落を含む論文の著者情報に基づき、同一研究者に関する情報を集約してオーソライズした形で提供する。具体的には、(1)書誌リンケージの結果や共著者などの多様な情報を手がかりにした同定候補の抽出、(2)機械学習による同定候補ペアの抽出、(3)グラフアルゴリズムの適用によるクリーニング、(4)識別子の付与と情報集約、の４つの手順を繰り返すことで、数千万人規模の著者同定を実現している。

従来の人物同定の研究では、検索クエリに対して得られる同姓同名人物のあいまい性解消に焦点があてられていたが、信頼性の高いサービスを実現するためには、表記は異なるが同一人物を参照する情報を発見し、同定することも同様に重要である。しかし表記揺れを許す場合には、検索の対象範囲が広がるため、同定精度および同定効率の両者で高い性能を実現することが要求される。提案手法では、すべての著者に異なる識別子を付与した状態から出発してボトムアップ的に処理を進め、書誌リンケージの結果、共著者情報、同一リスト中での共起、タイトルの類似性などの多様な文脈を使って同定候補を選別することで、この問題に対応している。また、ブートストラップ的な仕組みの導入により、表記揺れルールを自動的に追加して同一性の判定精度を高めたり、同定結果に基づきさらに新たな候補を獲得したりといった適応的な処理を実現し、さらに利用者からのフィードバックデータを取り込み矛盾なくデータを再構築するための仕組みも備えている。

本研究で提案した著者リンケージサーバの枠組みは、国立情報学研究所の学術コンテンツサービスである CiNii（http://ci.nii.ac.jp/）の実サービスにも取り込まれ、学術コンテンツのアクセス向上に貢献している。

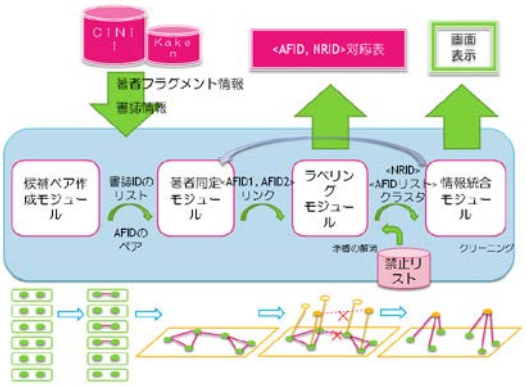

図 6　著者レインケージにおける処理の流れ

最後に、本研究の成果に基づき、論文検索結果を多様化するための推薦システムの実現法を検討し、研究者が過去に執筆した論文や論文どうしの引用関係に基づき多様な視点から論文を推薦するシステムを試作した。今後は、文書の内容解析を行うなどリンケージの対象を広げて、さらに高度な検索サービスの実現に向けて研究を進める予定である。

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 17 件）

① T. Akutsu, D. Fukagawa, A. Takasu: Approximating Tree Edit Distance through String Edit Distance, Algorithmica, 57(2), pp.325-364, 2010.

② H.Kurasawa, D.Fukagawa, A.Takasu, J.Adachi: Margin-based Pivot Selection for Similarity Search Indexes, IEICE Trans. Information and Systems, E93-D(6), pp. 1422-1432, 2010.
③ H.Kurasawa, A.Takasu, J.Adachi: Load Balancing Scheme on the basis of Huffman coding for P2P Information Retrieval, IEICE Trans. Information and Systems, E92-D(10), pp. 2064-2072, 2009.
④ T.Akutsu, D. Fukagawa, A.Takasu: Improved approximation of the largest common subtree of two unordered trees of bounded height, Information Processing Letters, 109(2), 165-170, 2008.
⑤ Q.M.Vu, A.Takasu, J.Adachi: Improving the Performance of Personal Name Disambiguation Using Web Directories, Information Processing & Management, 44(4), 1546-1561, 2008.
⑥ 相澤: 類語関係抽出タスクにおけるコーパス規模拡大の影響、情報処理学会論文誌, 49(3), pp.1426-1436, 2008.
⑦ Matsumura, A. Takasu, J. Adachi: Effect of Relationships between Words on Japanese Information Retrieval, ACM Trans. Asian Language Information Processing, 5(3), pp. 264-289, 2006.
.

〔学会発表〕(計 24 件)
① Q.M.,Vu, A.Takasu, J.Adachi: A Versatile Record Linkage Method by Term Matching Model Using CRF, Database and Expert Systems Applications (DEXA), pp. 547-562, 2009.
② H.Kurasawa, D.Fukagawa, A.Takasu, J.Adachi: Maximal Metric Margin Partitioning for Similarity Search Indexes, ACM Conf. Information and Knowledge Management (CIKM), pp.1887-1890, 2009.
③ D.Fukagawa, T.Akutsu, A.Takasu: Constant factor approximation of edit distance of bounded height unordered trees, String Processing and Information Retrieval Symp. (SPIRE), pp.7-17, 2009.
④ V.B.Dang, A.Aizawa: Multi-class named entity recognition via bootstrapping with dependency tree-based patterns, Pacfic-Asia Conf. Knowledge Discovery and Data Mining (PAKDD), pp.76-87, 2008.
⑤ Q.M.Vu, A.Takasu, J.Adachi: Name Disambiguation Boosted by Latent Topics from Web Directories, IEEE/WIC/ACM Intl. Conf. Web Intelligence (WI), 2008.
⑥ A.Takasu, D.Fukagawa, T.Akutsu: Statistical Learning Algorithm for Tree Similarity, IEEE Intl. Conf. Data Mining (ICDM), pp.667-672, 2007.
⑦ Y. Wang, K.Oyama: Framework for Building a High-Quality Web Page Collection Considering Page Group Structure, Asia-Pacific Conf. Web Conference (APWeb), pp.95-107, 2007.
⑧ Q.M.Vu, T.Masada, A.Takasu, J. Adachi: Using a Knowledge Base to Disambiguate Personal Name in Web Search Results, ACM Symp. Applied Computing (SAC), 2007.
⑨ M.Takaku,K.Oyama,A.Aizawa: An Analysis on Topic Features and Difficulties based on Web Navigational Retrieval Experiments, Asia Information Retrieval Symp.(AIRS), pp.625-632, 2006.
⑩ T.Akutsu, D.Fukagawa, A.Takasu: Approximating tree edit distance through string edit distance, Intl. Symp. Algorithms and Computation (ISAAC), pp.90-99, 2006.

６．研究組織
(1)研究代表者
安達　淳 (Adachi Jun)
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：80143551

(2)研究分担者
大山　敬三 (Oyama Keizo)
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：90177022

(3)連携研究者
高須　淳宏 (Oyama Keizo)
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

相澤　彰子 (Aizawa Akiko)
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：90222447

宮尾　祐介 (Miyao Yusuke)
国立情報学研究所・コンテンツ科学研究系・准教授
研究者番号：00343096