

## 能動的リソースマイニングに基づく異種情報統合基盤の研究

著者	北川 博之, 森嶋 厚行, 天笠 俊之, 川島 英之, 陳漢雄, 古瀬 一隆, 渡辺 陽介
発行年	2010-03
URL	<a href="http://hdl.handle.net/2241/105019">http://hdl.handle.net/2241/105019</a>

# 能動的リソースマイニングに基づく異種情報統合基盤の研究

研究代表者	北川 博之	筑波大学・大学院システム情報工学研究科・教授
研究分担者	森嶋 厚行	筑波大学・大学院図書館情報メディア研究科・准教授
	天笠 俊之	筑波大学・大学院システム情報工学研究科・准教授
	川島 英之	筑波大学・大学院システム情報工学研究科・講師
連携研究者	陳 漢雄	筑波大学・大学院システム情報工学研究科・講師
	古瀬 一隆	筑波大学・大学院システム情報工学研究科・講師
	渡辺 陽介	東京工業大学・学術国際情報センター・助教

## 1 研究の概要

ネットワーク上に分散した多様な情報を統合する情報統合について、これまで様々な研究がなされてきた。しかし、情報爆発の時代を迎え、情報源の数や規模の増加、情報源の異種性の増大、センサー等の動的な情報源の増大等により、有用な情報源を探索・発見し、適切な統合を図ることは一層困難となりつつある。よって、膨大かつ多様なネットワーク上の情報源の発見から統合にいたるプロセスをスケーラブルに実現する基盤技術が求められている。一方で、大量のデータから有用な知識を発掘するデータマイニングについても近年数多くの研究開発がなされているが、多くは個々の要素技術の開発にとどまっており、情報統合のプロセスとの融合を前提としたものではない。情報統合のプロセスに個々のデータマイニング技術を有機的に融合する包括的なフレームワークの構築は、大きな研究課題の一つである。本研究課題では、ネットワーク上に存在する多様なリソース（情報資源）を探索的にマイニングする技術を情報統合の枠組みに融合し、柔軟かつ拡張性のある情報統合を可能とするリソースマイニングに基づく異種情報統合基盤について研究開発を行っている。特に、情報源を発見する情報源マイニング、動的に変化する情報源を継続的にマイニングする連続的マイニング等の技術を開発すると共に、情報統合のベースとなる能動的情報統合基盤にこれらを融合することを目指す。具体的には、1) リソースマイニングを実現するためのマイニング要素技術、2) マイニングと情報統合に関わる応用研究、3) 情報統合基盤システムの研究開発、の3つの視点より研究を行った。

平成 21 年度は平成 18, 19, 20 年度の研究成果を踏まえて、1) に関しては、データストリームに対する外れ値検出、Web からの情報抽出、XML データに対する OLAP、高次元データに対する効率的な近傍検索等に関する研究成果を得た。2) に関しては、ソーシャルブックマークを用いた Web ページランキング、Web コンテンツ一貫性管理等に関する研究成果を得た。3) に関しては、分散ストリーム処理基盤における高信頼化方式、並列 XML 処理、セキュア情報統合、P2P 情報共有基盤の開発等を行った。

## 2 マイニングのための要素技術

### 2.1 データストリームに対する連続的外れ値検出手法

外れ値検出は、データマイニングの中でも重要な技術の一つであり、異常値検出等の様々な分野で応用されている。一方で、センサデバイスや金融市場等、データストリームを発信する情報源データが増えつつある。それゆえ、データストリームに対する外れ値検出の要求も増えつつある。データストリームに対する外れ値検出を実現するナイーブな方法は、静的なデータに対して用いられる手法をデータ到着毎に逐次適用することである。しかし、この手法は処理コストが高いため、頻繁に到着するデータストリームへの適用は効率面から好ましくない。

そこで本研究では、データストリームに対する効率的な連続的外れ値検出手法を開発した。昨年度までは、各時刻におけるデータ分布が直前の時刻のものと同様している場合が多いというデータストリームの特徴に着目して、差分計算による効率的な外れ値検出手法を開発した。そして同手法が直前時刻からの変化が大きくない場合において、セルベースである既存手法 [1] に比べて有用であることを示した。

本年度は昨年度の成果を受け、データストリームの変化量に応じて効率的な処理方法を選択するハイブリッド型手法を開発した。ハイブリッド型手法では、データストリームの変化量を表す「CDS<sup>1</sup>変化率」という概念を導入し、CDS 変化率が大きい場合は既存手法を、小さい場合は差分計算による手法を適用する (図 1)。実験により、ハイブリッド型手法の優位性を示した。

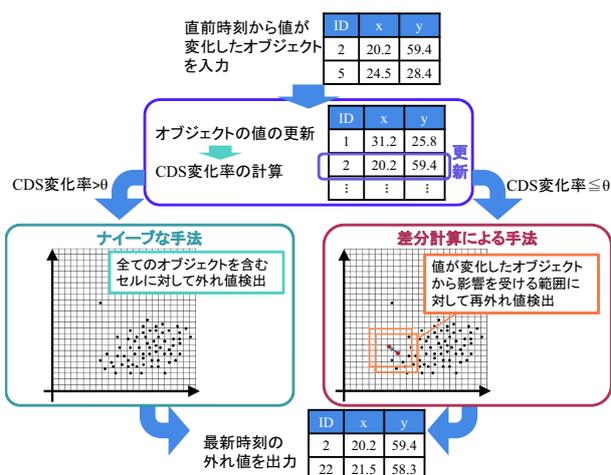


図 1: ハイブリッド型連続的外れ値検出手法

## 2.2 Web からの情報抽出

Web コンテンツなどの非構造情報からレコード情報を抽出する手法として情報抽出手法が注目され、この数年活発に研究されている。しかし、既存の手法は基本的に「(会社、所在地)」のような二項関係、あるいは単純な構造を持つレコード情報に特化しており、XML のような複雑な構造を抽出する手法はあまり議論されていない。そこで本研究では、既存の情報抽出手法を組み合わせることで、複雑な情報を抽出する手法を提案した。

種となる XML データからスキーマ情報を抽出し、XML データを関係表に写像する。写像した関係表を二項関係に分解することによって、DIPRE などの既存のレコード抽出手法が適用可能となる。得られた (単純な) レコードから、スキーマの構成情報を利用して XML データを再構成することにより、既存の情報抽出手法を利用しながら、より複雑な XML データの抽出が可能となる。実験により、提案手法の有効性を示した。

## 2.3 XML-OLAP

XML は標準データフォーマットとして広く用いられるようになり、膨大な情報リソースが日々 XML 形式で創出されている。XML で記述された情報源から、有益な情報を引き出すため、XML に対するマイニングが重要である。本研究では、マイニングの一手法である、対話的分析処理 (OLAP; Online Analytical Processing) を XML に適用した XML-OLAP に関する研究を行った。特に本年度は、XML データの階層構造の各レベルに対して異なる粒度の集約計算を行なう TOPOLOGICAL ROLLUP に着目し、それをマルチコア CPU 上で並列実行する手法を開発した。

前年度までの成果として、TOPOLOGICAL

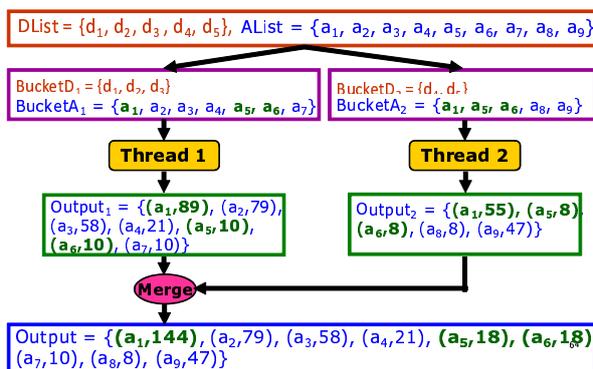


図 2: TOPOLOGICAL ROLLUP の並列処理

<sup>1</sup>Cell-based Data Stream

ROLLUP の処理において、XML データを一度走査するだけで、全てのレベルにおける集約結果を得ることができる SSC-Pre, SSC-Post を提案した。本年度は、このアルゴリズムを並列化するため、1) 入力データを木構造に基づき分割する方法、2) 並列版 SSC-Pre, SSC-Post アルゴリズム、3) 各スレッドで得られた部分計算結果を効率的にマージする手法を開発し、その有効性を実験により示した。

## 2.4 高次元データに対する効率的な $k$ -NN 検索手法及び画像検索への応用

高次元データは様々な分野で広く使われている。高次元空間を扱う場合、「次元の呪い (dimensionality curse)」と呼ばれる性質により、 $k$ -NN 検索性能が低下するという問題が起こる。このため、簡単な連続スキャン手法よりも、空間分割やデータ分割などの索引技術の性能が劣ることが指摘されている [6]。距離計算に  $L_p$  距離を用いる場合、 $p$  による冪乗計算のコストが非常に高いという問題が生じる [7]。

本研究では、 $k$ -NN 検索に対して距離関数に関する冪乗計算を格子に分割して索引を作成するという独自の関数索引手法を提案した (図 3 上)。提案手法では、高次元データを空間  $[0, 1]^D$  へ正規化し、座標空間を  $B$  ブロックに等分割する (例:  $B = 10$ )。それぞれの分割点は  $t/B$  となり、配列  $c[t] = (t/B)^p$  ( $t = 0, 1, \dots, B$ ) を求める。この配列と距離上界・下界の組合せにより、効率的な  $k$ -NN 検索が実現される。実データと合成データを用いた実験により、提案手法を連続スキャン法と R-tree 索引法と比較し、その優位性を示した。また、提案手法の有用性を示すために、画像検索システムを開発した (図 3 下)。このシステムでは収集した画像データに対して高次元特徴量の抽出と正規化を行った。

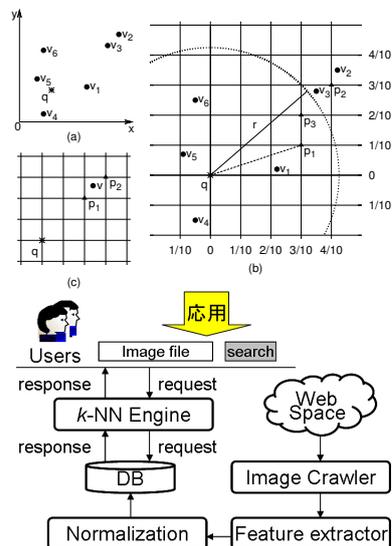


図 3: (上) 関数索引、(下) 画像検索

## 3 マイニングと情報統合に関する応用研究

### 3.1 ソーシャルブックマークの時系列変化を考慮した Web リランキング

ブックマーク情報にタグ付けし、Web 上で管理・分類・共有するソーシャルブックマーク (SBM) サービスが注目されている。SBM 上の Web ページは、ブックマークしたユーザにとって一定の価値のある情報と捉えることができる。そのため、被ブックマーク数はページの質を測る一つの指標と言えるが、その時間変化は従来研究では考慮されてこなかった。この点に着目し、我々はページの再ランキング手法 S-BITS を提案してきた。提案技法 S-BITS は、(ページ、ユーザ) 間のブックマーク関係から 2 部グラフを生成し、HITS [2] を適用することで、ページの Authority を測り Web ページのランキングを行う技法である。

本研究では、ブックマークの時系列情報に着目し、現時点におけるページコンテンツの価値の持続度を表す、鮮度の有無や活性度を推定した。また、鮮度の有無や活性度を加味した S-BITS の改良手法の提案を行った (図 4)。鮮度の有無は、ページの鮮度の有効期間を考慮し、直近のブックマークからの経過時間を基に推定した。活性度は、ページの平常時のブックマーク頻度を基準とし、ブックマーク頻度の時間変化から推定した。活性度推定法は、HMM を用いた Kleinberg のバースト検出法 [3]

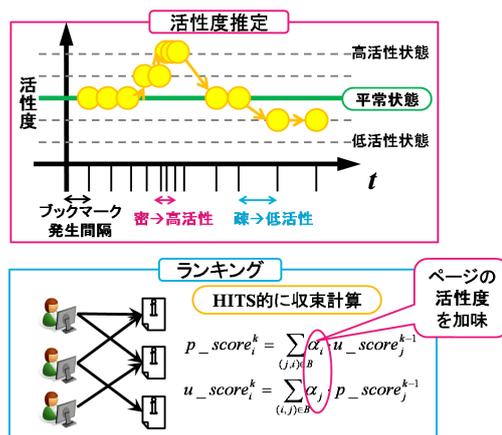


図 4: 活性度推定とランキング

に基づいた。これらの推定法は、ページ固有のブックマークの時系列情報から推定するアプローチを取っており、ページコンテンツの異種性を加味している。評価実験は、改良手法にある一定の有効性があることを示した。

### 3.2 Web コンテンツ一貫性管理

現在の Web アーキテクチャの特徴の一つはコンテンツの分散管理が容易であることであるが、その一方でコンテンツ一貫性の維持が難しいという問題がある。例えば、大学等では部局毎に Web サイトが存在し、また同一の Web サイトでもページ毎に管理者が異なっているということは良くあることであるが、その中に掲載された連絡先やイベント情報、論文リストなどに一貫性がないということがしばしば見受けられる。これらは利用者の混乱を引き起こすだけでなく、企業活動に関する Web サイトなどであった場合は、その組織に対しても不利益を引き起こす可能性もある。

一般に、Web コンテンツの一貫性管理を行うためには、バックエンドに DBMS を配置したシステムを構築し、DB に格納されたデータに基づいて Web ページを自動生成することによって関連するコンテンツの一貫性を維持するということが行われる。しかし、管理組織が異なる等の理由により、そのようなシステムを構築することが適切でない場合も多い。

本研究ではこれまで、分散管理された Web コンテンツの一貫性維持を支援する研究を推進してきた。本年度の成果は次の通りである。(1) 以前から推進してきた Web リンク切れ修正に関する成果発表 (2) 既存の Web コンテンツからコンテンツ一貫性制約を自動発見するアルゴリズムの効率化。また、本研究に関連して大量のリンク構造をデータベースで管理するためのグラフデータベースにおける効率的なデータ管理に関する研究についても推進した。

特に、今年度は (2) に関する研究を中心に推進した。具体的には、我々が提案する HTML 要素や XML 要素の内容を対象とした包含従属性 (Inclusion Dependencies) を効率よく発見する研究を行った。包含従属性はリレーショナルデータベースの分野でよく知られたデータ一貫性制約の一つであり、これまでも包含関係を手掛かりとした発見支援手法が研究されてきた。しかし、Web コンテンツを対象とした場合、論理的には包含関係があっても表記上の揺れやミスが存在するため、完全な包含関係を想定したアルゴリズムは適切ではない。我々は、完全ではないが程度の包含関係があると考えられる HTML・XML 要素の組合せを効率よく発見するため、ビットシグネチャを利用するアプローチの研究を行った。具体的には、各 HTML・XML 要素に対してビットシグネチャを計算し、包含関係が存在する可能性がない要素の組を排除することによって効率良い発見を支援する。実験の結果、本手法により 90%以上の組を候補から除去することができ、効率化に効果があることがわかった。

## 4 情報統合基盤

### 4.1 分散ストリーム処理基盤のための高信頼化方式

近年、センサーや IC タグ・カメラ等の実世界情報を提供するストリーム型の情報源を扱う情報統合・知識抽出の要求が高まってきている。我々はそのための基盤システムとして、StreamSpinner[4] の開発を行っている。StreamSpinner にはデータ到着やタイマーに連動し、イベント駆動的に各種統合処理を実行する機能が組み込まれており、利用者は SQL ライクな問合せ言語によって情報源に対する処理要求を与える事ができる。また、ユーザ定義の関数や演算を組み込むための仕組みを備えており、データマイニングの高度な知識抽出手法と連携した処理も可能である。また、地理的に分散した大量の情報源の利用や、特定ノードの負荷分散を実現させるために、

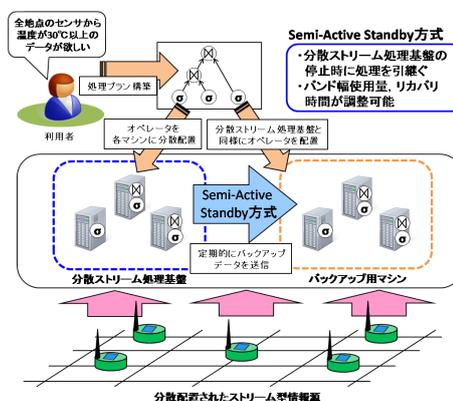


図 5: 分散ストリーム処理基盤の高信頼化

複数ノード上のストリーム処理基盤を協調動作させ得る分散ストリーム処理基盤の開発を行ってきている。

分散ストリーム処理基盤では、複数のノードが互いに連携し、情報源から利用者までストリームデータを中継していくことで処理を実現する。そのため、分散環境を構成する一部のノードが故障などにより停止してしまうと、ストリームデータは中継されず、システム全体が停止してしまう。更に、システムが停止している間もストリームデータは到着し続けるため、大量のストリームデータが失われることになる。このように分散ストリーム処理基盤では、ノードの停止によりシステム全体が停止しない、かつ、データの欠損が生じないという性質を満たす高信頼化方式が必要である。そこで、本研究では、分散ストリーム処理基盤のための高信頼化方式 Semi-Active Standby 方式 (以下 SAS 方式) を提案した (図 5)。SAS 方式では、高信頼化におけるバックアップデータの転送部分に「バッチ」という機構を導入し、データの送信制御を行う。従来の高信頼化方式 [5] とは異なり、利用者は「バッチ」を任意の値に設定することにより、要求に合わせて、バンド幅使用量とリカバリ時間の関係を調整することが可能である。評価実験により提案手法の優位性が示された。

## 4.2 並列 XML 処理

XML (Extensible Markup Language) は、データ表現のデファクトスタンダードとして広く利用されるようになった。このため、大量の XML データに対する高速な検索処理が求められている。一方、PC の低価格化と高性能化により、PC クラスタなどの計算機環境が身近なものとなった。さらに、一つの CPU に多数のコアを搭載したマルチコア CPU が一般化しており、このような環境を利用したデータベース検索の高速化に対するニーズは高い。

このような背景から、本論文では XML データに対する並列問合せ処理手法を提案する。特に、XML データの検索処理アルゴリズムである Holistic Twig Join を対象に、これを並列化する手法を開発した。Holistic Twig Join では、XML データがあらかじめ同名の要素が文書順に整列されたリスト (XML ノードストリーム) として格納されていることを仮定しているため、XML ノードストリームをオンラインで分割する手法を検討した。また、Holistic Twig Join の内部処理に基づいたステージングを行い、処理の効率化を図った (図 6。実験による評価によって、提案手法の有効性を示した。

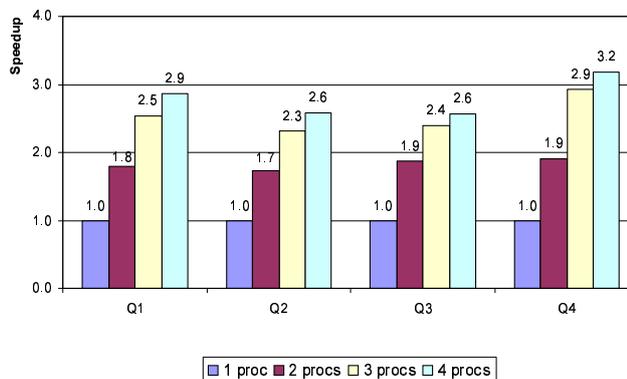


図 6: マルチコア CPU による Holistic Twig Join

このように背景から、本論文では XML データに対する並列問合せ処理手法を提案する。特に、XML データの検索処理アルゴリズムである Holistic Twig Join を対象に、これを並列化する手法を開発した。Holistic Twig Join では、XML データがあらかじめ同名の要素が文書順に整列されたリスト (XML ノードストリーム) として格納されていることを仮定しているため、XML ノードストリームをオンラインで分割する手法を検討した。また、Holistic Twig Join の内部処理に基づいたステージングを行い、処理の効率化を図った (図 6。実験による評価によって、提案手法の有効性を示した。

## 4.3 セキュア情報統合

セキュリティは、近年の Web アプリケーションにおいて重大な問題であることは周知のとおりである。特に、重要なビジネスデータは攻撃者の格好のターゲットとなっている。このため、データの機密性、プライバシー、一貫性を保証することは、データベースシステムにおいてきわめて重要な課題である。特に最近注目を集めた情報漏えい事件を見ると、外部からの防御だけでは不十分であり、内部か

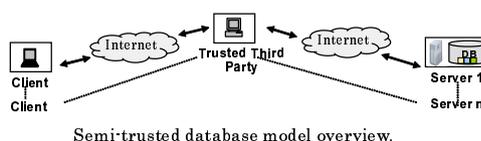


図 7: MCDB の概要

らの攻撃を防ぐ手だても必要である。このためには、データベースの暗号化が強力な手段となるが、単一の鍵で暗号化されたデータベースは、いったん鍵が攻撃者の手に渡ってしまうと、データベース全体を盗まれてしまうという危険性がある。

このような背景から、本研究では、インターネットのようなオープンな環境で、企業や部署などのいくつかの主体がデータを交換する環境を想定し、複数の暗号化手法を組み合わせることでデータの機密性、プライバシー、一貫性を保証する MCDB (Mixed Cryptography Database) を提案した。この手法では、データベースのカラムを異なる鍵で暗号化しておく。問合せの際には、信頼できる第三者機関 (Trusted Third Party) を通じて問合せおよび問合せ結果を適切に暗号化、書き換えることによって、複数の主体にまたがった問合せを安全に実行することができる (図 7)。

#### 4.4 P2P 情報共有基盤

我々は新しいスケーラブルな分散索引構造 RCAN を提案した。同構造は動的 P2P システムにおける多次元データの管理を対象とする。RCAN は多環 CAN<sup>2</sup> である。RCAN は純粋な P2P オーバレイであり、完全分散化と自己組織化を実現している。また、RCAN において、各ノードはルーティング状態を自己調整する。RCAN には 2 つの特徴がある。まず、RCAN は効率的なルーティングを実現している。N ノードから構成されるネットワークにおいて、各ノードは  $O(\log N)$  本の長距離リンクを有する。完全照合と範囲問合せは  $O(\log N + \log M)$  ホップでルーティングされる。ここで  $M$  は問合せ範囲と交差するノード数を表す。次に、RCAN は負荷均一化を実現している。具体的には、RCAN は複数結合方式と系列結合方式を組み合わせたハイブリッド方式を提供する。シミュレーション実験により、提案方式が従来方式に対して優れることを確かめた。

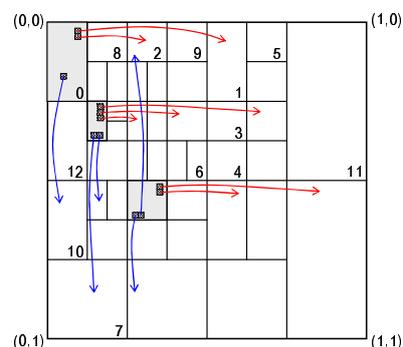


図 8: Multi-Ring Content Addressable Network

### 5 まとめと今後の展開

以上述べたように、平成 18, 19, 20 年度の研究成果を踏まえて、本年度は、1) リソースマイニングを実現するためのマイニング要素技術、2) マイニングと情報統合に関わる応用研究、3) 情報統合基盤システムの研究開発、の 3 つの視点より研究開発を推進し、各研究項目に関して所定の成果を得た。特に、今年度の課題としてきた情報抽出型マイニング技術、メタ情報の分析・マイニング技術、メタデータ型以外の情報統合環境等についても、本報告書に述べたような形で一定の研究の進展を行うことができた。今後は、それぞれの研究項目に関する技術の深化や一般化をさらに図りたい。また、最終年度を迎えるに当たり、これまでに研究開発した要素技術の融合の方向を探りたい。特に、マイニング要素技術と情報統合技術の融合については、適切な対象を絞り込んだ上で研究を推進していきたい。さらに、これまでも進めてきた InTrigger を利用した情報統合基盤システムの研究開発についても一層の進展を図りたい。

### 研究成果リスト

#### 著書、論文

1. Hiroyuki Kitagawa, Yousuke Watanabe, Hideyuki Kawashima, and Toshiyuki Amagasa: “Stream-based Real World Information Integration Framework”, Wireless Sensor Network Technologies for Information Explosion Era (Springer book series “Studies in Computational Intelligence”), Springer Verlag, (to appear).

<sup>2</sup>Multi-Ring Content Addressable Network

2. Atsuyuki Morishima, Akiyoshi Nakamizo, Toshinari Iida, Shigeo Sugimoto and Hiroyuki Kitagawa: “Why are Moved Web Pages Difficult to Find? The WISH Approach”, Proc. 18th Int. World Wide Web Conference, pp.1117–1118, 2009.
3. Hasan Kadhem, Toshiyuki Amagasa and Hiroyuki Kitagawa: “Encryption over Semi-trusted Database”, Proc. DASFAA2009 PhD Workshop, LNCS5667, pp.358–362, 2009.
4. Tsubasa Takahashi and Hiroyuki Kitagawa: “A Ranking Method for Web Search Using Social Bookmarks”, Proc. International Conference on Database Systems for Advanced Applications (DASFAA 2009), pp.585–589, 2009.
5. Hasan Kadhem, Toshiyuki Amagasa and Hiroyuki Kitagawa: “A Novel Framework for Database Security based on Mixed Cryptography”, Proc. International Conference on Internet and Web Applications and Services(ICIW 2009), pp.163–170, 2009.
6. Chantola Kit, Toshiyuki Amagasa and Hiroyuki Kitagawa: “Algorithms for Structure-based Grouping in XML-OLAP”, International Journal of Web and Information Systems, Vol. 5, No. 2, pp.122–150, 2009.
7. Imam Machdi, Toshiyuki Amagasa and Hiroyuki Kitagawa: “XML Data Partitioning Schemes for Parallel Holistic Twig Joins”, International Journal of Web and Information Systems, Vol. 5, No. 2, pp.151–194, 2009.
8. Atsuyuki Morishima, Akiyoshi Nakamizo, Toshinari Iida, Shigeo Sugimoto and Hiroyuki Kitagawa: “Bringing Your Dead Links Back to Life: A Comprehensive Approach and Lessons Learned”, Proc. the 20th ACM Conference on Hypertext and Multimedia (ACM Hypertext 2009), pp.15–24, 2009.
9. 寺島慎太郎、天笠俊之、北川博之: “木直列化に基づく XML データの類似結合における木構造の統合”, 日本データベース学会論文誌, Vol. 8, No. 1, pp.47–52, 2009.
10. Djelloul Boukhelef and Hiroyuki Kitagawa: “Efficient Multidimensional Data Management in Structured Peer-to-Peer Systems”, PhD Workshop (VLDB 2009), 2009.
11. 大喜恒甫、渡辺陽介、北川博之、川島英之: “対象情報源を動的に選択可能なストリーム処理機能の実装と評価”, 情報処理学会論文誌: データベース (TOD43), Vol. 2, No. 3, pp.1–17, 2009.
12. MoonBae Song and Hiroyuki Kitagawa: “Managing Frequent Updates in R-trees for Update-intensive Applications”, IEEE Transactions on Knowledge and Data Engineering, Vol. 21, No. 11, pp.1573–1589, 2009.
13. Imam Machdi, Toshiyuki Amagasa and Hiroyuki Kitagawa: “Executing Parallel TwigStack Algorithm on a Multi-core System”, Proc. 11th International Conference on Information Integration and Web-based Applications and Services (iiWAS2009), pp.14–16, 2009.
14. Hasan Kadhem, Toshiyuki Amagasa and Hiroyuki Kitagawa: “Mixed Encryption over Semi-Trusted Database”, MASAUM Journal of Basic and Applied Science(MJBAS), Vol. 1, No. 2, pp.302–312, 2009.
15. Djelloul Boukhelef and Hiroyuki Kitagawa: “Efficient Load Balancing Techniques for Self-organizing Content Addressable Networks”, Special issue on Selected Paper of ICUIMC 2009, Journal of Networks, (to appear).

## 受賞

1. 高橋公海、森嶋厚行、杉本重雄、北川博之: “Web ページを対象とした包含従属性の効率的な発見手法”, 第1回データ工学と情報マネジメントに関するフォーラム (DEIM2009) 優秀インタラクティブ賞 受賞, 2009.

## その他

1. 北川博之: “大規模センサデータ処理のためのデータストリーム管理基盤”, イノベーション・ジャパン 2009 大学見本市, 2009 年 9 月 16–18 日 (展示会および新技術説明会).

## 参考文献

- [1] E. M. Knorr, R. T. Ng and V. Tucakov, “Distance-based Outliers: Algorithms and Applications”, VLDB Journal, Vol. 8, Issue 3-4, pp. 237–253, 2000.
- [2] J. Kleinberg. “Authoritative Sources in a Hyperlinked Environment” Journal of the ACM, Vol. 46, No. 5, pp. 604–632, 1999.
- [3] J. Kleinberg. “Bursty and hierarchical structure in streams”, Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 91–101, 2002.
- [4] StreamSpinner project, <http://www.streamspinner.org/>.
- [5] J. H. Hwang, M. Balazinska, A. Rasin, U. Cetintemel, M. Stonebraker, and S. Zdonik, “High-Availability Algorithms for Distributed Stream Processing”, Proc. IEEE International Conference on Data Engineering, pp. 779–790, 2005.
- [6] S. Berchtold, C. Böhm, D. Keim, and H.-P. Kriegel. “A Cost Model for Nearest Neighbor Search in High-Dimensional Data Space”, Proc. ACM Symposium on Principles of Database Systems (PODS), pages 78–86, 1997.
- [7] C. Aggarwal, A. Hinneburg, and Daniel A. Keim. “On the Surprising Behavior of Distance Metrics in High Dimensional Spaces”, Proc. International Conference on Data Theory, pp. 420–434, 2001.