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/計算科学研究センター・教授
研究分担者　石川 佳治　名古屋大学・情報連携基盤センター・教授
天笠 俊之　筑波大学・大学院システム情報工学研究科/計算科学研究センター・講師
森嶋 厚行　筑波大学・図書館情報メディア研究科・助教授

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．初年度である平成１８年度は，１）に関しては，比率規則マイニング，XML データに対する OLAP，話題構造マイニング，人物の呼称情報抽出，時系列文書クラスタリング，移動統計情報抽出等に関する研究成果を得た．２）に関しては，Web 連続モニタリングによるページ移動先探索と DB と連携した文書情報源からの情報抽出に関する研究成果を得た．３）に関しては，拡張性を有する能動的情報統合基盤システムのプロトタイプを開発した．

2　マイニング要素技術に関する研究

2.1　比率規則マイニング

本研究では，数値データに対するデータマイニング技術を扱っている．数値データの相関性抽出の一つとして，データがもつ線形関係を表す比率規則を抽出する手法がある [1]．比率規則は欠損値の補完，予測，外れ値検出など多種の応用が可能であるため，その抽出は重要な技術課題である．

既存の比率規則抽出手法として代表的な，主成分分析を用いた手法 [1] では，複数の線形関係が混在するような状況において，個々の線形関係を捉えられないことがある．また線形関係を直線として捉えるため，部分的に成り立つ線形関係を捉えることができない．

これに対し本研究では，比率規則を 2 次元空間中の線分とその周辺領域のタプルが満たす性質と定義し，領域内に含まれているタプル (個々のデータ) はその比率規則に従うと定義した．さらに，全データ中に対する割合を表すサポートと，属性値がある区間に含まれるタプル中で比率規則に従う割合を表す確信度を

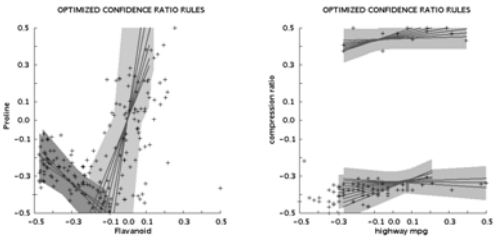

図 1: 実データに適用し得られた比率規則の例

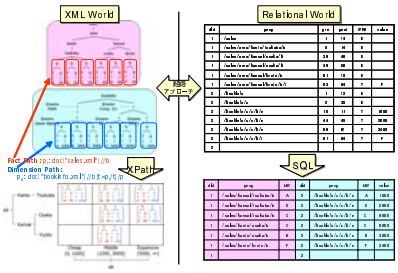

図 2: XML データに対する OLAP の全体構造

定義した．既存の比率規則抽出手法では与えられたデータに対して得られるマイニング結果は固定的であるが，本研究ではサポートと確信度を用いることで，ユーザの意図を反映した結果を得ることが可能である．また，拡張手法として，クラスタリングと組み合わせた手法を提案した．この手法ではタプルの分布が偏った場合にもその影響を軽減し，妥当な線形関係を捉えることができる (図 1)．

今後の課題としては，より幅広い種類の実データに対する検証，高次元データへの適用手法の検討，得られた比率規則の応用等が挙げられる．

2.2 XML データに対する OLAP

XML (Extensible Markup Language) は，データフォーマットの事実上の標準として広く認知され，情報統合のための基盤となるデータ表現・交換手段として用いられている．今後，XML 形式で蓄積されるデータ量は爆発的に増加することが予想されるため，大量の XML データから有益な情報をマイニングするための分析処理が重要となる [2]．そこで本研究では，複雑かつ対話的な XML データの分析処理を可能にする XML-OLAP (Online Analytical Processing) 技術の研究開発を進めている．

平成１８年度は，提案手法の基礎的な部分に主眼を置いて研究を進めた．具体的には，１）XML データ上の多次元キューブの形式的定義，２）リレーショナル DBS を利用した XML キューブ上の OLAP 演算の実現方式の 2 点を検討した．XML データは木構造を有し，一般に分析対象のデータ，分析データに付随する関連データ，分析と関係しないデータが混在している．この中から多次元データ分析を行うための事実データと次元データを抽出するため，本研究では XPath 式を用いる．ファクトを抽出するための XPath 式を Fact Path と呼び，次元を抽出するための XPath 式 Dimension Path は Fact Path の相対パスとして与えられる．これによって抽出される部分 XML データを各次元値でグループ分けし，集約演算を適用することで XML データの多次元分析が可能となる (図 2 左)．特に，XML データでは従来の数値属性以外に，XML データそのものが持つ入れ子構造によるグルーピングを考慮することが重要である．本研究ではそのために各部分 XML データの絶対パス式を利用したグルーピング操作を提案した．

さらに，以上の操作を実現するためリレーショナル DBS を利用した実装方式を提案した (図 2 右)．基本的には，XML データをリレーショナルデータベースに格納するための経路アプローチ [3] に基づいており，任意の XML データ上の多次元キューブと分析処理を関係データベースの機能だけを用いて実現することができる．予備実験による評価を行った結果，数百 MB 程度のサイズの XML データに対して，実用的な速度で分析処理が可能であることを示した．

今後は，XML データに内在する「データ指向」の部分と「文書指向」の差異に着目し,「文書指向」XML データの特性を分析処理に取り込むための処理機構を検討する．また，XML スキーマ語彙のオントロジやパス式の類似度を考慮したマッチング処理を検討する．

2.3 話題構造マイニング

“文書” は重要な情報源であり，大規模な文書集合からの情報抽出は，情報統合を実現する上で欠くことのできない技術である．本研究では，文書集合中の主要な話題が知りたい，文書集合中の特定の話題に関する文書を参照したい，といった要求に対応するための技術を研究開発している．

このような要求に対する従来の解決手段として，クラスタリングがある．しかし，クラスタリングには，閾値設定や，クラスタ割り当て以上の情報が得られないといった問題点がある．

これらの問題に対応するため，本研究では以下に示すステップから構成される話題構造マイニングを提案している．

1. 文書集合中の文書をノード，文書間の関係をエッジで表現したグラフ構造を構築
2. グラフ構造に基づき各ノードの中心性 (PageRank 値) を算出
3. グラフ構造および中心性スコアに基づき話題構造の抽出

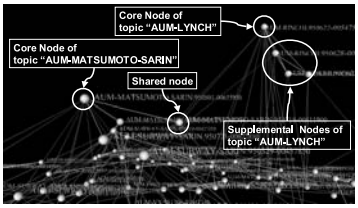

図 3: 話題構造マイニングにより得られた話題構造

上記により，文書を話題毎に分類するだけでなく，文書集合中に存在する主要な話題を特定し，複数の話題間の関係や，話題の中心的な内容に対する文書の位置付けを明らかにする事が可能となる (図 3)．

このように文書もしくは言語的な要素の集合に対して，要素間の類似性を元にしたグラフ構造を考慮する手法は Graph-based NLP と呼ばれており，文書のランキングや，重要文抽出 [4] に利用されている．ただし，従来手法は，ノードの中心性を元にしたランキングもしくはランキング上位の情報の選択を行っていたのに対し，提案手法ではノードの中心性とグラフ構造の両方を利用し，ノード間もしくはノード群間の関係を抽出する点が従来研究と異なる．

今後は文書間類似度以外の尺度の導入，文書以外のテキスト要素の利用，計算量低減といった方向で，提案手法の適用範囲の拡大を行う予定である．

2.4 人物の呼称情報抽出

様々な情報源に分散する情報を統合する上で，同一オブジェクトの同定は極めて重要である．このためのアプローチとして，本研究では人物オブジェクトの同定の着目し，同一人物に対する呼び名 (呼称) を Web から抽出する手法を提案した．提案手法は，１）文字列 *alias* が，*fullname* という名前の人物の呼称であることを述べる際，日本語では “*alias* こと *fullname*” と表現する場合が多い，２）人物のフルネームと呼称は，同様な局所的コンテクスト中に出現することが多い，という２つのヒューリスティクスを用いる．具体的には，以下の３ステップで人物の呼称を抽出する．

1. “こと *fullname*” という文字列を含む Web 文書群より，対象人物の呼称候補を抽出
2. “*fullname*” を含む Web 文書群より，フルネームと隣接する文字列 (prefix, suffix パターン) を抽出・重みづけを行い，重要なパターンを「隣接パターン」として選択
3. 抽出した呼称候補と隣接パターンを接続させて Web 検索クエリを作成，各クエリに対応する「(推定) 検索結果数」とパターンの重みを用いて呼称候補を評価し，評価値の高い上位 k 件の呼称候補を，人物の呼称として抽出

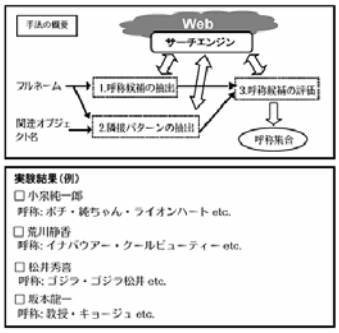

図 4: 呼称の抽出手法と実験結果

実際の Web データを用いた実験により，小泉純一郎 (純ちゃん，ライオンハート等)，荒川静香 (イナバウアー，クールビューティー等)，松井秀喜 (ゴジラ，ゴジラ松井等) など複数の有名人について妥当な呼称が抽出できることを確認した (図 4)．

今後は人物がフルネームでなく，非公式な呼び名 (呼称) で参照されている Web ページや Blog 記事を検索・特定する．また，それら特定したページ (記事) 集合から，対象人物に関する評判情報，口コミ情報など，動的/非公式な情報・知識を抽出することを検討する予定である．

2.5 時系列文書クラスタリング

インターネット上では，大量のテキスト情報が日々流通しているが，あまりにその量が膨大であるため，有効な利用のためには情報の集約技術が求められる．特に本研究では，連続的マイニングに関する技術の一つとして，ニュース記事などのように，発行された日時などの情報が文書に付随するような時系列的な文書の集約技術に焦点を当てている．文書の**クラスタリング**技術は，情報検索の分野を含め古くから研究されているが，本研究では特に**新規性に基づくクラスタリング** (novelty-based clustering) というアプローチを提案している (図 5)．

図 5: 提案するクラスタリングシステム

ニュース記事などでは新しい情報がより重視され，古い情報は時間が経つにつれて忘却されていくが，時系列的な文書に対するこのような価値の概念を文書類似度の導出に用いる．これにより，古い情報は積極的に忘れ去られ，結果として新しい情報を中心としてクラスタリングが行われるようなクラスタリングを実現する．また，オンライン環境では，時々刻々と配信される文書に対してクラスタリングを効率的に実現する必要がある．そこで本研究では，k-means 法に拡張を行い，インクリメンタルな更新処理を実現している．Topic Detection and Tracking (TDT) プロジェクト [7] のように，時間的な文書に着目してトピックの検出や追跡を行う研究はこれまでも見られたが，新規なトピックを中心にクラスタリングして提供するというアプローチやインクリメンタルな更新を工夫している点が本研究の特徴である．

今後は，本クラスタリング手法に連携したユーザインタフェース技術の開発や，Blog などの情報源への適用などについても検討を進める．

2.6 移動体統計情報抽出

GPS による位置測位技術の発達や無線ネットワークの普及などにより，近年では，多数の**移動オブジェクト** [6] をリアルタイムに追跡することが容易となっている．このような背景を受け，本研究では，連続的マイニングに関する技術の一つとして，多くの移動オブジェクトの移動軌跡をコンパクトに集約する**移動ヒストグラム** (mobility histogram) の構築手法に関する研究を行った (図 6)．

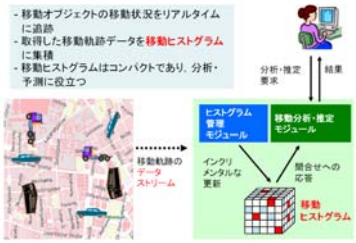

図 6: 移動体統計情報抽出の概要

本研究の特徴は，グリッド状に分割した空間内でのマルコフ連鎖により移動パターンを表現する点にあり，移動ヒストグラムはマルコフ連鎖に基づく各パターンの出現頻度を蓄積する．論理的なヒストグラムはデータキューブの形式で表現されるが，物理的には木構造を用いてコンパクトなデータ構造を実現する．

移動オブジェクトを対象とした場合，各移動オブジェクトからストリーム状に連続的に送信されてくる移動軌跡の情報をリアルタイムに集約可能であることが重要である．そのため，本研究ではストリーム的な移動軌跡データに対して低コストで更新可能なデータ構造および更新処

理のアルゴリズムを開発した．また，移動ヒストグラムにはそのサイズがコンパクトであることが求められることから，与えられたサイズの上限のもとで高精度のヒストグラムの実現も進めた．

今後，より柔軟な移動パターンの表現方式や，移動ヒストグラムを利用した移動情報のマイニングなどの技術について展開することを検討している．

3　マイニングと情報統合に関わる応用研究

3.1　Web 連続的モニタリングによるページ移動先探索

Web 情報の統合利用における問題の一つに，分散して存在するコンテンツやそれらの間の一貫性維持が困難なことがある．その一例として，リンク切れの問題は以前から重要な問題であると認識されている [9]．そこで本研究では Web リンクの一貫性維持の問題に着目し，Web ページのリンク切れが発見されるとページの移動先の探索を行うことによって，リンクの一貫性維持を支援するシステムの研究を行った．

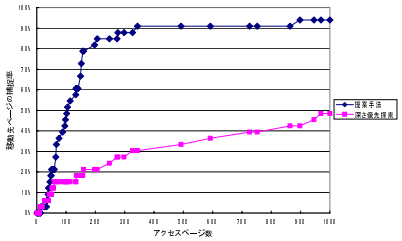

図 7: 同一サイト内における移動先ページの捕捉率

本研究の特徴は次の通りである．すなわち，他の手法 [10] では，移動先ページがあらかじめサーチエンジン等にインデックスされていることを前提としているのに対し，本研究での手法では，移動先ページがインデックスされていなくても「移動先ページが存在しそうな場所」を効率よくクロールすることによって，ページの移動先を高い発見率で発見することである．具体的には，ロボットを用いて Web を連続的にモニタリングした結果からページの移動先でありそうな場所を計算し，リンク切れの際にはその情報を用いて効率よく移動先ページの発見を行う．

本研究に関しては次のような成果が得られた．１）開発した手法を実装したプロトタイプシステムを構築した．２）本プロトタイプシステムを用いて，実 Web データを対象とした大規模実験を行った．３）本プロトタイプを一般利用可能な環境の構築を行い Web サイトにて公開した．４）過去の実験結果から同一サイト内での Web ページの移動が多いことが明らかになったため，同一サイト内での移動ページの探索において高い探索効率を実現する手法の開発を行った．本手法を用いることにより，移動先ページの発見に必要なページアクセス数の削減が実現可能なことがわかった (図 7)

今後は，１）ページの移動先発見のための計算コストのより詳細な検証，２）リンク切れ以外の Web コンテンツ一貫性維持の問題，について取り組んでいきたい．

3.2　DB と連携した文書情報源からの情報抽出

Web 上には膨大な情報が存在し，かつ拡大を続けている．しかし，Web 上の情報源の異種性等により，膨大な情報の有効な利用は必ずしもなされていないというのが現状である．Web や各種文書情報源から有用な情報を得るための手法として，情報抽出に関する研究がこれまで行われてきたが，膨大な情報源から，ユーザが必要とする情報を選択的に抽出するかが重要な問題となっている．

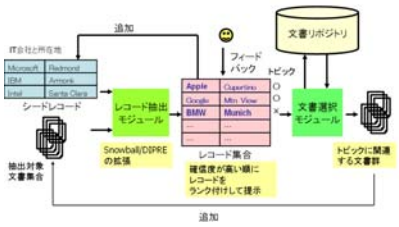

図 8: 情報抽出手法の概略

このような背景に基づき，本研究では，情報源マイニングの考え方を導入した，データベースと連携した文書情報源からの情報抽出手法の開発を進めている (図 8)．従来の情報抽出では，文書データのみが主に利用されたのに対し，本研究では既存のデータベースの情報を情報抽出手法と連携し，情報抽出における精度と抽出数の向上に役立てる点に一つの特徴がある．また，ユーザからの例示レコードおよび提示されたサンプルデータベースの情報をもとに，関連の深いレコードデータを中心に抽出する点にも特徴がある．従来の情報抽出手法では，例

示レコードのトピックを考慮せず，やみくもに多くのレコードを抽出しようとしていたのに対し，選択的な抽出を目指している．情報抽出方式としては，ブートストラップ型 [8] と呼ばれる情報抽出の方式を基盤としており，スケーラブルな情報抽出を目指している．

今後は，データベース管理システムの問合せ処理機能を有効活用する，より高度な連携技術の開発や，データクリーニング手法と連携したノイズの低減手法の開発などの実現を図る．

4　情報統合基盤システムの研究開発

本研究では，各種データマイニング機能を組み込んだ能動的情報統合を実現するための基盤システムの研究開発を行っている．従来型のデータベース等の静的な情報源のみでなく，センサー等の動的な情報源を含めた情報統合を実現するため，本システムにはデータ到着やタイマーに連動し，イベント駆動で能動的に各種統合処理を実行する機能が組み込まれている．また，利用者は SQL ライクな問合せ言語によって各種統合要求を与える事ができる．本研究では，特に，１）各種マイニング手法との連携のための拡張性，２）センサーネットワーク等の分散環境に対応した能動型情報源統合環境の実現についての研究開発を行っている．

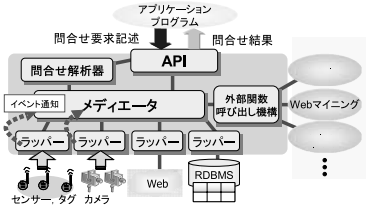

図 9: 情報統合基盤システム

１）については，これまで我々が研究開発を進めてきた基盤システムにおける問合せ言語ではリレーショナル代数の範囲内の要求であれば記述可能であった．しかし，本研究課題が目指す，より高度な情報統合を実現するためには，各種データマイニング手法との連携が必須となる．本研究では，これまでに開発してきた基盤システムに外部のプログラムを呼び出す仕組みを取り入れ，問合せから外部関数として呼び出す機能を実現した．具体的には，カメラからの映像ストリームに対する意味情報として，顔認識ライブラリを呼び出し，映像に写っている人物の ID を付加し，それを情報統合に用いるといった使い方が可能である．現在の実装では，Java で開発されたプログラムであれば，基盤システム本体のコードに一切手を加えることなく容易に組み込む事ができる．今後は，この仕組みを利用して，様々なマイニング手法を組み込む実験を行う予定である．

２）については，センサーネットワーク等を含めた分散ネットワーク上で情報統合基盤システムを協調動作させる，分散・能動型情報統合環境の開発を行った．開発した基盤システムでは，例えば，情報源に近い位置のノードで一旦フィルタリングや集約を行ってデータ量を減らし，その後利用者に近い位置にあるノードへ配信して他のノードからの情報や他の情報源と統合する，といったことが可能である．現状では利用者が自ら各ノードに割り当てる処理を指定することが必要だが，今後は，システム側で各ノードの負荷やネットワーク帯域を考慮した処理の割り振りを行うことを検討している．

5　今後の展望

以上述べたように，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所定の成果を得た．今後は，上記に記載した各研究テーマに関して残された課題に取組むと共に，各種マイニング要素技術を情報統合の枠組みに融合するための方式に関する検討を具体化したいと考えている．また，支援班 UCN グループによるセンサーネットワークテストベッド環境への本研究における情報統合基盤システムの適用についても検討を進めたい．

参考文献

[1] F. Korn, A. Labrinidis, Y. Kotidis, and C. Faloutsos, ”Quantifiable Data Mining Using Ratio Rules”, VLDB Journal, 8(3-4): 254-266, 2000.

[2] R. R. Bordawekar and C. A. Lang： “Analytical Processing of XML Documents: Opportunities and Challenges”, ACM Sigmod Record, Vol. 34, No. 2, pp. 27–32, 2005.

[3] M. Yoshikawa, T. Amagasa, T. Shimura, and S. Uemura: “XRel: A Path-Based Approach to Storage and Retrieval of XML Documents using Relational Databases”, ACM Transactions on Internet Technology, Vol. 1, No. 1, pp. 110-141, June 2001.

[4] R. Mihalcea, and P. Tarau: “TextRank: Bringing Order into Texts,” Proc. 2004 Conference on Empirical Methods in Natural Language Processing, pp.404–411, Barcelona, Spain, July 2004.

[5] S.Tejada et al., Learning Domain-Independent String Transformation Weights for High Accuracy Object Identification, In SIGKDD 2002.

[6] Güting, R.H. and Schneider, M.: “Moving Object Databases”, Morgan Kaufmann, 2005.

[7] Allan, J. (ed.): “Topic Detection and Tracking: Event-based Information Organization”, Kluwer, 2002.

[8] Agichtein, E. and Gravano, L.: “Querying Text Databases for Efficient Information Extraction”, Proc. ICDE, pp. 113-124, 2003.

[9] Robert P. Dellavalle, Eric J. Hester, Lauren F. Heilig, Amanda L. Drake, Jeff W. Kuntzman, Marla Graber, Lisa M. Schilling. “Going, Going, Gone: Lost Internet References”, Science 302: 787-788

[10] Park, S.-T., Pennock, D. M., Giles, C. L., and Krovetz, R. 2004. “Analysis of lexical signatures for improving information persistence on the world wide web”, *ACM Trans. Inf. Syst. 22,* 4, 540–572.

研究成果リスト

著書，論文

1. 濱本雅史，北川博之: “サポートと確信度をもとにした比率規則による線形関係抽出”, 情報処理学会論文誌：データベース, Vol. 47, No. SIG19(TOD32), 2006(採録決定).
2. 戸田博之，北川博之，藤村考，片岡良治，奥雅博: “グラフ分析を利用した文書集合からの話題構造マイニング”, 電子情報通信学会論文誌, Vol. J90-D, No. 2, 2007(採録決定).
3. 石川佳治，町田陽二，北川博之: “マルコフ連鎖モデルに基づく移動ヒストグラムの動的構築法”, 電子情報通信学会論文誌, 2007(採録決定).
4. 町田 陽二, 石川 佳治, 北川 博之: “マルコフ連鎖モデルに基づく動的な移動ヒストグラム構築手法”, 日本データベース学会 Letters, Vol. 5, No. 1, pp.89–92, 2006.
5. Sophoin Khy, Yoshiharu Ishikawa, and Hiroyuki Kitagawa: “Incremental Clustering Based on Novelty of On-line Documents”, 日本データベース学会 Letters, Vol. 5, No. 1, pp.57–60, 2006.
6. 外間智子, 北川博之: “Web データを用いた人物の呼称抽出”, 日本データベース学会 Letters, Vol. 5, No. 2, pp.49–52, 2006.
7. Sophoin Khy, Yoshiharu Ishikawa, and Hiroyuki Kitagawa: “Novelty-based Incremental Document Clustering for On-line Documents”, Proceedings of International Workshop on Challenges in Web Information Retrieval and Integration (WIRI 2006), 2006.
8. Yoshiharu Ishikawa, Yoji Machida, and Hiroyuki Kitagawa: “A Dynamic Mobility Histogram Construction Method Based on Markov Chains”, Proceedings of 18th International Conference on Scientific and Statistical Database Management (SSDBM 2006), pp.359–368, 2006.
9. Masafumi Hamamoto and Hiroyuki Kitagawa: “Ratio Rule Mining with Support and Confidence Factors”, Proceedings of 3rd IEEE International Conference on Intelligent Systems (IS 2006), pp.500–505, 2006.
10. Hiroyuki Toda, Ryoji Kataoka, and Hiroyuki Kitagawa: “Topic Structure Mining for Document Sets using Graph-Based Analysis”, Proceedings of 17th International Conference on Database and Expert Systems Applications (DEXA2006), pp.327–337, 2006.
11. Tomoko Hokama and Hiroyuki Kitagawa: “Extracting Mnemonic Names of People from the Web”, Proceedings of 9th International Conference on Asian Digital Libraries (ICADL 2006), pp.121–130, 2006.
12. Hiroyuki Toda, Ko Fujimura, Ryoji Kataoka, and Hiroyuki Kitagawa: “Topic Structure Mining using PageRank without Hyperlinks”, Proceedings of 9th International Conference on Asian Digital Libraries (ICADL2006), pp.151–162, 2006.

13. Shinichi Yamada, Yousuke Watanabe, Hiroyuki Kitagawa and Toshiyuki Amagasa: “Location-based Information Delivery Using Stream Processing Engine StreamSpinner”, Proceedings of Intl. Conf. on Mobile Data Management, pp.57, 2006.

14. Tomoko Hokama and Hiroyuki Kitagawa: “Detecting ”Hot” Topics about a Person from Blogspace”, Proceedings of 16th European - Japanese Conference on Information Modeling and Knowledge Bases (EJC 2006) , pp.290–294, 2006.

15. 濱本雅史，北川博之: “局所性を考慮した比率規則マイニング”, 電子情報通信学会技術研究報告 Vol. 106, No. 150, pp.25–30, 2006.

16. 外間智子, 北川博之: “Web コーパスを用いた人物の呼称抽出”, 電子情報通信学会技術研究報告 Vol. 106, No. 149, pp.145–152, 2006.

17. Sophoin Khy, Yoshiharu Ishikawa, and Hiroyuki Kitagawa: “Parameter Setting for a Clustering Method through an Analytical Study of Real Data”, 電子情報通信学会技術研究報告, Vol. 106, No. 150, pp.43–48, 2006.

18. Chantola Kit, Toshiyuki Amagasa, Hiroyuki Kitagawa: “Towards Analytical Processing of XML Data”, 電子情報通信学会技術研究報告 Vol. 106, No. 148, pp.163–168, 2006.

19. 張建偉, 黒川沙弓, 石川佳治, 北川博之: “フィードバックを利用した文書の選択に基づくレコード抽出手法”, 電子情報通信学会技術研究報告, Vol. 106, No. 149, pp.227–232, 2006.

20. 石川佳治, 黒川沙弓, 張建偉, 北川博之: “データクリーニングを統合した情報抽出システムの提案”, 電子情報通信学会技術研究報告, Vol. 106, No. 150, pp.61–66, 2006.

21. 渡辺陽介, 山田真一, 北川博之: “分散環境におけるストリーム処理の高信頼化”, 電子情報通信学会技術研究報告, Vol. 106, No. 149, pp.203–208, 2006.

22. 山田真一，渡辺陽介，北川博之: “ストリーム管理システムにおける永続化要求の妥当性評価”, 電子情報通信学会技術研究報告, Vol. 106, No. 149, pp.209–214, 2006.

23. 石川佳治: “マルコフ遷移モデルに基づく移動軌跡のクラスタリング”, 電子情報通信学会技術研究報告, Vol. 106, No. 290, pp.37–42, DE2006–125, 2006.

公開ソフトウェア

1. 森嶋厚行，飯田敏成，澤菜津美，中溝昌佳，有山智洋，杉本重雄，北川博之: PageChaser: An Automatic Web-link management Tool developed by the WISH Project

   http://wish.slis.tsukuba.ac.jp/LIM-RO.html

   Web リンクの自動監視およびリンク切れ時のページ移動先発見プログラム.

受賞

1. 濱本雅史: ” 局所性を考慮した比率規則マイニング”, 夏のデータベースワークショップ DBWS2006 学生発表奨励賞, 2006.

2. Kit Chantola: ”Towards Analytical Processing of XML Data”, 夏のデータベースワークショップ DBWS2006 学生発表奨励賞, 2006.

3. 外間智子: ”Web コーパスを用いた人物の呼称抽出”, 夏のデータベースワークショップ DBWS2006 学生発表奨励賞, 2006.