

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・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
陳 漢雄 筑波大学・大学院システム情報工学研究科・講師

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ます、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 19 年度は平成 18 年度の研究成果を踏まえて、1) に関しては、比率規則マイニング、外れ値検出、XML データに対する OLAP、時系列文書クラスタリング、集約最近傍検索等に関する研究成果を得た。2) に関しては、文書情報源からの情報抽出、XML データの類似検索、Web コンテンツ一貫性管理、Web 情報源分類等に関する研究成果を得た。3) に関しては、拡張性と能動性を有する分散型情報統合基盤システムのプロトタイプを開発した。

2 マイニングのための要素技術

2.1 比率規則マイニング

比率規則 [1] は数値属性間における線形関係を表したものであり、欠損値の補完、予測、外れ値検出など多様な応用が可能である。主成分分析等の既存の比率規則抽出では、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、部分的に成り立つ線形関係を捉えることができない。

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はその比率規則に従うと定義する。線分を考慮するため、線分の傾きだけでなくデータ中の比率規則の位置に応じて様々な比率規則を抽出することが可能である。また、相関規則マイニングにおける場合と

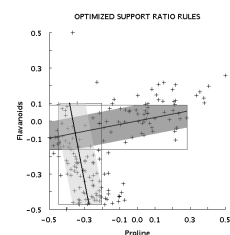


図 1: 比率規則の抽出。

同様に、サポートと確信度を定義し、利用者の意図を反映した比率規則を得ることができるようになった(図1)。本年度は、昨年度開発した非対称比率規則が有する問題を解消する対称比率規則を開発すると共に、高次元空間への拡張の可能性に関する検討を行った。

2.2 外れ値検出

2.2.1 高次元データセットにおける外れ値検出

外れ値検出は、データから興味ある知見を得るデータマイニング手法の一つである。従来の外れ値検出手法は、高次元データに対しては次元の呪いにより有効に機能しないことが知られている。また、既存のアルゴリズムは、最適な性能を得るために、事前に幾つかのパラメタを設定する必要があるが、最適なパラメタ選択は容易ではない。

本研究では、例示に基づく高次元データに対する外れ値検出アプローチを開発した。例示に基づく外れ値検出方法は、利用者がどのような外れ値に興味があるかを得る上で有望である。提案手法の概要を図2に示す。提案手法では、元の高次元空間そのものではなく例示外れ値を最も良く特長付ける低次元の部分空間を探す。その探索には、遺伝子アルゴリズムを用いる。得られた部分空間において、例示外れ値と同様の振舞いをするオブジェクトを外れ値として検出する。本研究のもう一つの重要なポイントは、利用者により提供された例示にノイズが含まれ得ることを考慮する点にある。このため、低次元部分空間の探索においては、例示外れ値の多くが外れ値として振舞うものを探索することとしている。

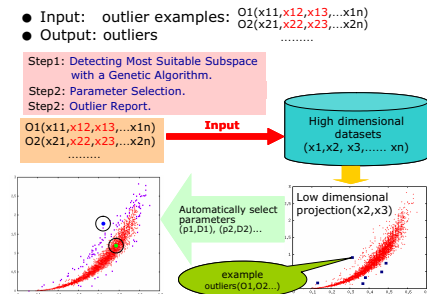


図 2: 高次元データセットにおける外れ値検出。

2.2.2 トランザクションデータからの外れ値トランザクション検出

既存の外れ値検出手法の多くは、数値データを対象としている。トランザクションデータ等、カテゴリカルなデータに対する外れ値検出は、これまであまり研究がなされていない。

本研究では、トランザクションデータにおけるアイテム間の相関性に着目し、トランザクション中のアイテム集合と本来ならば共起するはずのアイテムが出現していないようなトランザクションを外れ値とみなし検出する手法を開発した。トランザクションの外れ値度は、高い確信度を持つ相関ルールを用いて、そのトランザクション中に本来出現するはずのアイテムを数え上げることで計算する。また、ブルートフォースアルゴリズムは計算コストが非常に大きいため、外れ値度の計算に影響しない冗長な相関ルールの除去を行い、外れ値度の上限值を求め、候補外れ値トランザクションを絞り込むことで、より高速な外れ値トランザクション検出を実現した。

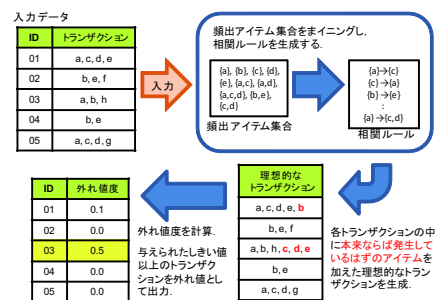


図 3: トランザクションデータに対する外れ値検出。

実データを用いた実験により、次のような成果が得られた。(i) 提案された外れ値度による検出で、比較対象である確率計算を利用した関連研究 [2] を上回る精度を得られた。(ii) ブルートフォースアルゴリズムと比較して、提案アルゴリズムがより効率的であることを確認した。

2.3 XML-OLAPによるXMLデータの対話的分析処理

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。

本年度は、昨年度に続きXML-OLAPの基礎部分の研究を進めるとともに、XMLキューブの計算を高速化するための手法について検討した。提案するXML-OLAPでは、分析の対象となるファクト(事実)とファクトを分類するためのディメンション(次元)をパス形式のXML問合せとして与える。抽出されたファクトをディメンションの値によって集約計算することで、多次元的な視点からの分析が可能になる。以上の機能を実現するため、本研究ではリレーショナルDBMSを利用した実装方式を提案している(図4)。基本的には、XMLデータを経路アプローチに基づきデータベースし、任意のXMLデータ上の多次元キューブと分析処理をデータベース機能だけを用いて実現することができる。

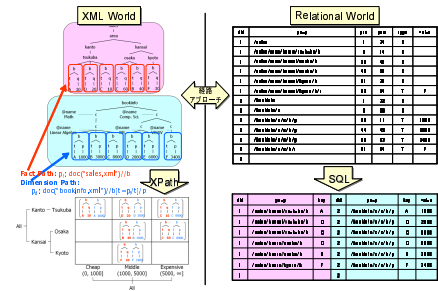


図4: XML-OLAPシステム。

2.4 時系列クラスタリング

インターネット上のニュースサービスなどの普及により、今日ではネットワークを介して大量の文書データがユーザに配信されている。時々刻々と配信される莫大な情報から必要な情報を抽出する労力は多大であるため、有効な利用のためには情報の集約技術が求められる。

本研究では、ニュース記事等のように、発行日時などの情報が付加された時系列文書の集約技術に焦点を当て、特に新規性に基づくクラスタリングというアプローチを提案している(図5)。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。基本的なアイデアは、文書の時間的な忘却の概念を導入し、文書が古くなるほどその価値が減少するというモデル化を行い、文書類似度を導出する点にある。これにより、新規性の高い文書を中心にクラスタリングを行うことが可能となる。また、*K*-means法を拡張したインクリメンタルな更新処理を実現し、評価実験により有効性を評価した。

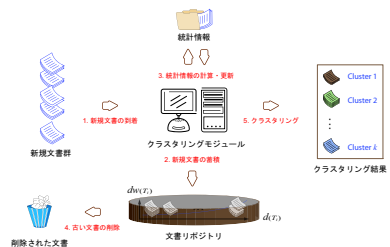


図5: 文書クラスタリングシステム。

2.5 集約最近傍検索のための索引手法

本研究は、多次元データの索引・効率的な検索手法の研究の一環として近傍検索(Nearest Neighbor)の一般化問題を取り扱う。Aggregate Nearest Neighbor問題は複数の問合せ点が与えられるとして、これらの問合せ点に対して特定の合計関数(aggregate distance)値を距離として評価し、最小距離のデータ点を求める。図6(左)では三つの問合せ点(q_1, q_2, q_3)に対して、合計距離あるいは最大距離が最小なるデータ点 p_x を求めている。このとき、合計関数がsum(max)ならば、答えは $p_5(p_7)$ になる。

この問題に対して、いままででは索引機構が適用できないと考えられており、[3]で提案されたアルゴリズムが有効であるとされてきた。この方法は、問合せ点の一つ一つについて、その最近傍の点を順次調べ、合計関数値が最小になるかどうか調べなければならない。本研究では距離に基づく索引が有効であることを発見し、図6(右)に示されたように、探索範囲を狭めるアルゴリズムを開発した。具体的には、三角不

等式に基づくフィルタリングを複数回行うことで、限られたデータだけを探索することでより効率的な処理が可能になる。

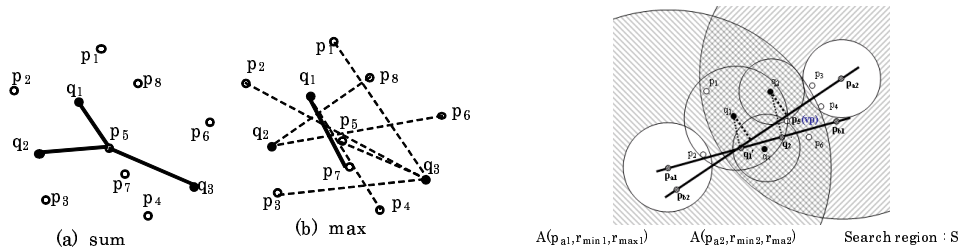


図 6: 集約最近傍探索 (左) / 提案手法 (右) .

3 マイニングと情報統合に関する応用研究

3.1 例示レコードに基づく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拡大を続けている。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行われてきたが、膨大な情報源から利用者が必要とする情報を選択的に抽出することは容易でない。

本研究では、情報源マイニングの考え方を導入した、例示レコードに基づく文書情報源からの情報抽出手法の開発を進めている (図 7)。従来の情報抽出手法 (例えば [4]) 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、提案手法は選択的な抽出を行なうことを目的に、適切なレコード抽出用情報源を求めることに焦点を当てている。

選択される文書情報源を優先的に処理することで、処理コストを削減すると同時に、それらの文書から利用者が必要とするレコードを高い抽出精度で取得することができる。本年度は、五つの文書選択手法を実装し、それらを比較することで提案手法の改良や有効性評価を行った。

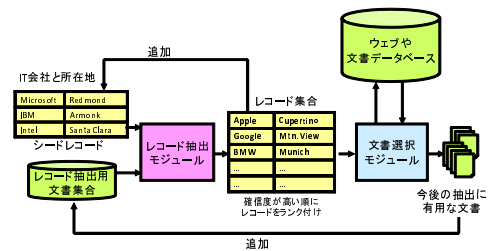


図 7: 情報抽出手法の概要 .

3.2 XML データの類似検索

XML で記述されるデータは急激に増加しており、ネットワークを通じて膨大な XML 形式の情報源へのアクセスが可能となっている。XPath や XQuery 等の従来型の問合せ言語は、パス表現を基礎としているため、XML データの構造を予め知っておく必要がある。しかし、今後は構造を知らないデータに対しても検索や統合等の処理手段を提供することが重要になると考えられる。

本研究では、次の二つの項目について研究を行なった: 1) オントロジを用いた XPath 問合せ書き換えによる XML データのあいまい検索、2) 木直列化を用いた XML データの類似結合。1) については、構造は知らないが内容は知っているという

ことを前提に、内容が類似しかつ構造も知っている XML データに対する問合せをまず利用者にかかせ、それをオントロジを用いて書き換える。これにより、構造を知らない XML データに対する問合せを得ることができる。2) については、類似した内容を持つ XML データ同士の類似結合を行なうため、まず XML データを直列化し、適切な長さの部分系列に変換する。得られた系列同士の類似度を図ることにより、類似

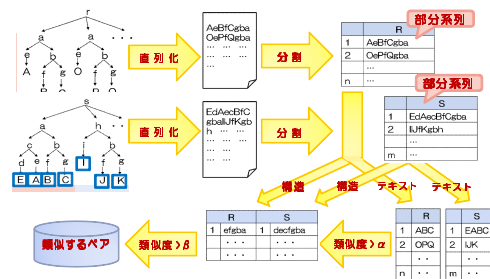


図 8: XML データの類似結合 .

した内容を持つ XML 部分木のペアを得ることができる (図 8)。これらの手法について、実験による評価を行ない、それぞれが有効であることを示した。

3.3 連続モニタリングによる Web コンテンツ一貫性管理

現在の Web アーキテクチャの問題の一つに、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。本年度は、これまで取り組んできた Web リンク一貫性維持の問題に加えて、リンク以外の Web コンテンツの一貫性維持のための基盤技術についても研究を推進した。

本研究の特徴は次の通りである。すなわち、既存の Web コンテンツの一貫性管理では、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り、関連するコンテンツの一貫性を維持するということが一般的であった。しかし、このアプローチではシステム構築コストが高くなるといった問題や、既存 Web コンテンツのボトムアップ的な統合管理が困難であること、非定型コンテンツへの対応に向いていないこと、と

といった問題がある。それに対し、本研究で開発中の手法では、バックエンド DB の存在を前提とせず、直接 Web コンテンツ間の制約を明示的に指定し、それを利用して一貫性の管理を実現するというアプローチを取る (図 9)。本年度は、特に Web コンテンツに存在する制約を効果的に記述するための技術基盤として、Web コンテンツのラッピング技術について研究を推進した。これは既存の比較的大がかりな手法 [5, 6] と異なり、一貫性制約の維持に必要な部分だけを容易にラッピングすることを目的としている。

本年度は次のような成果が得られた。(1) Web リンク一貫性維持の問題に関しては、開発した手法を実装したシステムを用いて実験を進め、詳細な分析を行った。(2) Web コンテンツ一貫性維持の問題に関しては、軽量ラッピング言語 Wraplet を開発した。

3.4 Web 情報源分類

本研究では、リンク解析手法を応用することにより、Web ディレクトリを自動拡張する手法の構築を行った。Web ディレクトリは検索エンジンと並んで情報取得手段として有効であるが、サイトの分類を手動により行うことから、情報爆発時代の時代にあっては情報源として網羅性に欠ける傾向にあるという欠点がある。そこで本研究では、未登録の Web サイトを階層構造を持った既存の Web ディレクトリに分類するものである。リンク解析を用いることで、Web サイトの記述言語に依存せず、画像等の非テキストが主体となる対象でも対応可能である。

本手法では、新たな Web サイトの分類のためにディレクトリ従属値という指標を導入している。ディレクトリ従属値は、検索エンジン等で利用されている PageRank と同様の考え方によって信頼性の高いリンク情報を基に既存 Web ディレクトリ内の Web サイトの評価を行ない、その評価を関連サイト検索アルゴリズムにおいてリンクの重みとして利用するものである。また、基本アルゴリズムを拡張して、トップディレクトリから再帰的に評価を実行する改良版も開発し、より分類精度の向上を行っている。本年度は、考案した拡張手法を既設 Web ディレクトリに適用することで、その精度を検証し、分類精度の向上を確認した。

4 情報統合基盤システム

近年、センサーや IC タグ・カメラ等のストリームを含めた多種多様な情報源が利用可能となっており、それらに対する情報統合要求が高まってきている。そこで、我々は情報統合のための基盤システムとして、StreamSpinner の開発を行っている。ストリームを含めた情報統合を実現するため、StreamSpinner には

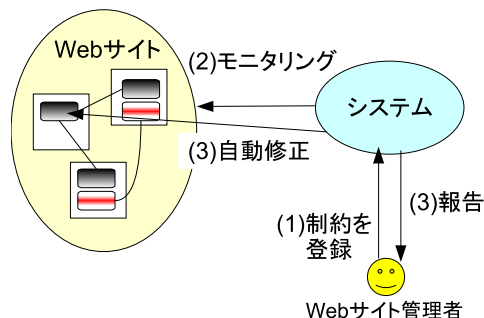


図 9: 明示的なコンテンツ一貫性制約を用いた Web サイト管理手法。

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おり、利用者はSQLライクな問合せ言語によってストリームに対する問合せ要求を与える事ができる。また、データマイニングなどの知識抽出手法との連携を想定し、ユーザ定義関数による任意の処理を追加できるようになっており、高い拡張性を備えている。本年度は、特に、(i) 大規模カメラ映像ストリームからの知識抽出、(ii) 広域分散した情報源に対する情報統合のための問合せ最適化手法の研究を行った。

(i) について、従来の映像解析システムは映像単体の処理に特化し、異種情報源の統合に関してはほとんど考慮がされていなかったが、本研究はカメラ映像ストリームと他の情報源との統合に重点を置く。まず StreamSpinner の拡張性を生かして、既存の映像解析手法を StreamSpinner 上のユーザ定義関数として実装した。これにより、例えば、映像ストリーム中の画像同士の変化を検出し、変化量が一定値以上の場合にだけ、他の情報源との統合処理を実行する、といった要求が容易に記述できる。また、映像解析処理は通常のリレーショナル演算処理に比べると負荷が大きく、リソースに応じて同時に処理対象にできるカメラ数が限られるため、選択的な情報源の利用が求められる。例えば、大量のカメラの中から特定人物の写った映像を抽出するようなアプリケーションでは、人物の位置情報を用いて、その人物の近傍のカメラ映像だけを処理対象とするといったことが必要となる。そこで本研究では、利用者が問合せ要求の中でそのような動的情報源選択を指定するための枠組みを導入した。今後は、シミュレータなどを用いた評価実験により、大規模カメラ環境における提案手法の有効性を検証する予定である。

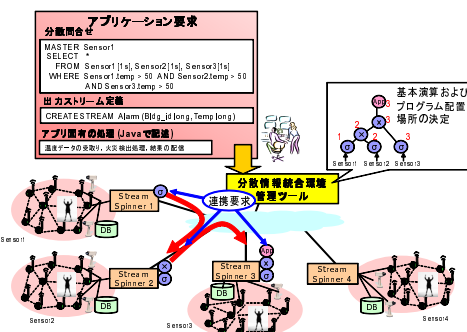


図 10: 分散・能動型情報統合環境。

(ii) についてはこれまで、地理的に分散した大量の情報源を扱うために、複数ノード上の StreamSpinner を協調動作させる、分散・能動型情報統合環境の実装を行ってきた (図 10)。提案システムは、利用者に情報源の分散を意識させずに、情報統合アプリケーションの開発が行える環境を提供することを目的としている。そのため、提案システムは、利用者のプログラムを分散環境中のどのノードに配置し、どのような経路で分散環境から必要なデータを収集すれば効率が良いかを利用者の代わりに判断する必要がある。本年度は、そのための分散問合せ最適化手法の提案を行った。分散問合せ最適化手法 [7] はすでにいくつか提案されているが、これらは分散環境内で通信遅延やデータレートの変化が起きた際に、演算を再配置することによって性能の低下を防ぐというものである。しかし、提案システムでは利用者プログラムの配置も含めて考慮する必要があり、内部で任意の処理を行っている可能性のあるプログラムの再配置は安易には実行できない。提案最適化手法は、再配置可能な処理と再配置不可能な処理が混在した状況において、通信遅延やデータレートの変化に対して、効率低下を防げるような配置場所を決定する。現在、情報爆発プロジェクトの基盤システムである InTrigger 上で、提案システムを用いた評価実験を実施中である。

5 まとめと今後の展開

以上述べたように、平成 18 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。それぞれの研究項目に関しては、手法の有効性のさらなる検証、手法の一般化、効率の処理等の面で残された課題があり、今後は、これらに取り組んでいく予定である。また、マイニング技術により検出したある種のイベントに連動して情報統合を行う機構に関しても今年度の成果を踏まえて、より研究を進展させたい。さらに、情報統合基盤システムを InTrigger 上で動作させる実験を進めているが、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 J., 8(3-4):254-266, 2000.
- [2] K. Das and J. G. Schneider, Detecting anomalous records in categorical datasets, Proc. KDD, pp. 220-229, 2007.
- [3] D. Papadias, Y. Tao, K. Mouratidis, and C. K. Hui, Aggregate Nearest Neighbor Queries in Spatial Databases. ACM TODS. 30(2):529-576, 2005.
- [4] E. Agichtein and L. Gravano, Snowball: Extracting relations from large plain-text collections”, Proc. ACM DL, pp. 85–94, 2000.
- [5] L. Liu, C. Pu, and W. Han, XWRAP: An XML-enabled Wrapper Construction System for Web Information Sources, Proc. ICDE, pp. 611-621, 2000.
- [6] A. Arasu, H. Garcia-Molina, Extracting Structured Data from Web Pages, Proc. SIGMOD, pp.337-348, 2003.
- [7] P. R. Pietzuch et al., Network-Aware Operator Placement for Stream-Processing Systems, Proc. ICDE, p. 49, 2006.

研究成果リスト

著書、論文

1. 澤菜津美, 森嶋厚行, 飯田敏成, 杉本重雄, 北川博之: “Web ページ移動先発見のための効率的なクローリング手法”, 情報処理学会論文誌: データベース, Vol. 48, No. SIG 11 (TOD 34), pp.149-161, 2007.
2. 濱本雅史, 北川博之: “対称比率規則の抽出手法”, 日本データベース学会 Letters, Vol. 6, No. 1, pp.73-76, 2007.
3. 石川雅弘, 吉川昂伯, 陳漢雄, 古瀬一隆, 大保信夫: “類似部分検索のためのタイムワーピング距離の下限值計算”, 日本データベース学会 Letters, Vol. 6, No. 1, pp.25-28, 2007.
4. 張建偉, 石川佳治, 北川博之: “トピックを考慮した大規模文書情報源からのレコード抽出”, 情報処理学会論文誌: データベース, Vol. 48, No. SIG 14 (TOD 35), pp.107-123, 2007.
5. Chantola Kit, Toshiyuki Amagasa, and Hiroyuki Kitagawa: “OLAP Query Processing for XML Data in RDBMS”, 3rd IEEE Int'l Workshop on Databases for Next-Generation Researchers (SWOD'07), 2007.
6. Yuan Li and Hiroyuki Kitagawa: “DB-Outlier Detection by Example in High Dimensional Datasets”, 3rd IEEE Int'l Workshop on Databases for Next-Generation Researchers (SWOD'07), 2007.
7. Jianwei Zhang, Yoshiharu Ishikawa, and Hiroyuki Kitagawa: “Record Extraction Based on User Feedback and Document Selection”, Joint Conf. of 9th Asia-Pacific Web Conference and 8th Int'l Conf. on Web-Age Information Management (APWeb/WAIM'07), pp.574-585, 2007.
8. Yanmin Luo, Hanxiong Chen, Kazutaka Furuse, and Nobuo Ohbo: “Finding Aggregate Nearest Neighbor Efficiently without Indexing”, INFOSCALE'07, 2007.
9. Masafumi Hamamoto and Hiroyuki Kitagawa: “Locality-aware Ratio Rule Mining”, 4th Int'l Conf. on Fuzzy Systems and Knowledge Discovery (FSKD'07), pp.693-698, 2007.
10. Hanxiong Chen, Kousuke Yamamoto, Kazutaka Furuse, and Nobuo Ohbo: “Supporting Web Search with Near Keywords”, 4th Int'l Conf. on Fuzzy Systems and Knowledge Discovery (FSKD'07), pp.411-416, 2007.

11. Yanmin Luo, Hanxiong Chen, Kazutaka Furuse, and Nobuo Ohbo: "Efficient Methods in Finding Aggregate Nearest Neighbor by Projection-based Filtering", Int'l Conf. on Computer Science and Applications (ICCSA'07), pp.821-833, 2007.
12. Toshiyuki Amagasa, Lianzi Wen, and Hiroyuki Kitagawa: "Proximity Search of XML Data using Ontology and XPath Edit Similarity", 18th Int'l Conf. on Database and Expert Systems Applications (DEXA'07), pp.298-307, 2007.
13. Natsumi Sawa, Atsuyuki Morishima, Shigeo Sugimoto, and Hiroyuki Kitagawa: "Wraplet: Wrapping Your Web Contents with a Lightweight Language", 3rd IEEE Int'l Conf. on Signal-Image Technology and Internet-based Systems (SITIS'7), pp.16-19, 2007.
14. Hiroyuki Toda, Hiroyuki Kitagawa, Ko Fujimura, and Ryoji Kataoka: "Topic Structure Mining using Temporal Co-occurrence", 2nd Int'l Conf. on Ubiquitous Information Management and Communication (ICUIMC'08), 2008 (to appear).
15. Yousuke Watanabe, Ryo Akiyama, Kousuke Ohki and Hiroyuki Kitagawa: "A Video Stream Management System for Heterogeneous Information Integration Environments", 2nd Int'l Conf. on Ubiquitous Information Management and Communication (ICUIMC'08), 2008 (to appear).
16. Yuan Li and Hiroyuki Kitagawa: "Example-Based Robust DB-Outlier Detection for High Dimensional Data", 13th Int'l Conf. on Database Systems for Advanced Applications (DASFAA'08), 2008 (to appear).
17. Lianzi Wen, Toshiyuki Amagasa, and Hiroyuki Kitagawa: "An Approach for XML Similarity Join using Tree Serialization", 13th Int'l Conf. on Database Systems for Advanced Applications (DASFAA'08), 2008 (to appear).
18. Atsuyuki Morishima, Akiyoshi Nakamizo, Toshinari Iida, Shigeo Sugimoto, Hiroyuki Kitagawa: "PageChaser: A Tool for the Automatic Correction of Broken Web Links", 24th Int'l Conf. on Data Engineering (ICDE'08), 2008 (to appear).
19. Kazuyo Narita and Hiroyuki Kitagawa: "Detecting Outliers in Categorical Record Databases Based on Attribute Associations", 10th Asia-Pacific Web Conf. (APWeb'08), 2008 (to appear).

受賞

1. 戸田浩之, 北川博之, 藤村考, 片岡良治: "時間的近さを考慮した話題構造マイニング", 電子情報通信学会第18回データ工学ワークショップ (DEWS'07) 優秀論文賞受賞, 2007.
2. 濱本雅史, 北川博之: "対称比率規則の抽出手法", 夏のデータベースワークショップ 2007 学生発表奨励賞受賞, 2007.
3. 成田和世, 北川博之: "トランザクションデータベースに対する高確信度の相関ルールを用いた外れ値検出手法", 夏のデータベースワークショップ 2007 学生発表奨励賞受賞, 2007.
4. Sophoin Khy, Yoshiharu Ishikawa, Hiroyuki Kitagawa: "A Document Clustering Method Focusing on User Specified Time of Interest", 夏のデータベースワークショップ 2007 学生発表奨励賞受賞, 2007.
5. Yuan Li, Hiroyuki Kitagawa: "Detecting Outliers in High Dimensional Datasets with Examples", 夏のデータベースワークショップ 2007 学生発表奨励賞受賞, 2007.
6. 澤菜津美, 森嶋厚行, 杉本重雄, 北川博之: "非定型 Web コンテンツ管理のための軽量ラッピング言語", 夏のデータベースワークショップ 2007 学生発表奨励賞受賞, 2007.