様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２５年３月３１日現在

機関番号：２５４０３
研究種目：基盤研究（Ｃ）
研究期間：２００８年度～２０１２年度
課題番号：２０５００１３７
研究課題名（和文）　知識創造支援型データベースシステムの構成法と効率化に関する研究

研究課題名（英文）　Study on Organizing a Knowledge-Creating Database and Achieving Efficient Processing

研究代表者
北上　始（KITAKAMI HAJIME）
広島市立大学・情報科学研究科・教授
研究者番号：５０２３４２４０

研究成果の概要（和文）：
知識創造支援型データベースの帰納的な仕組みについては、ギブスサンプリングや類似文字列検索の結果として得られるミスマッチクラスタから規則性を見つけ出す方法を明らかにした後、マルチコア PC クラスタ上でその方法を並列化することに成功した。知識創造支援型データベースの演繹的な仕組みについては、空間的な座標配列データに対する類似構造検索の精度を向上する方法を明らかにした。知識創造支援型データベースの発想的な仕組みについては、さまざまなコミュニケーションの基礎的事項について検討した。

研究成果の概要（英文）：
In the research related to inductive mechanisms for knowledge-creating databases, the problem of extracting regularity from a mismatch cluster that was defined as a set of similar subsequences returned by either Gibbs sampling or approximate query processing has been solved, where the regularity was represented as a set of minimum generalized patterns with a regular expression. Moreover, parallelization has been achieved for extracting the regularity from the mismatch cluster using Multicore PC clusters. In the authors' research about deductive mechanisms for knowledge creating-databases, the problems have been solved to increase accuracy of approximate query processing in coordinate sequence databases. In the authors' research related to abduction mechanisms for knowledge-creating database, various communications were considered.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2008 年度	900,000	270,000	1,170,000
2009 年度	600,000	180,000	780,000
2010 年度	600,000	180,000	780,000
2011 年度	700,000	210,000	910,000
2012 年度	800,000	240,000	1,040,000
総　計	3,600,000	1,080,000	4,680,000

研究分野：総合領域
科研費の分科・細目：情報学・知能情報学
キーワード：テキストマイニング、配列マイニング、索引構造、正規表現、並列処理

１．研究開始当初の背景

グローバル化、サービス部門の著しい増加、インターネットの驚異的な普及、 ビジネスの動的なコンポーネント化にともない、サー

ビス指向アーキテクチャや Web サービスといった研究開発分野が注目されている。これは、2006 年 7 月に米国の計算機科学分野の学会誌 CACM でも紹介されたが、この研究開発分野の出現によって、情報システムの時代から経済的利益によって動機づけられるサービスシステムの時代への移行が予測された。このような予測を受け、現在は、サービスシステムがもつ資産価値や知識資産の創造という観点から従来の情報システムとくにデータベースを内蔵する情報システムをさらに強化あるいはまったく新しい知識創造支援システムを構築しなければならない時代に突入しているともいえる。

本研究では、人間の知識創造プロセスを演繹・帰納・発想から成るという視点（哲学者パースが提唱）に「記憶」という視点を追加し、資産価値や知識資産の創造の支援をめざした知識創造型データベースシステムの研究を行う。従来のデータベース技術は演繹やオブジェクトなどのコンセプトを含んでいたが帰納や発想というコンセプトが含まれていない。このため、人間の知識創造過程の支援という点から評価すると極めて力不足である。我々は、従来のデータベース技術に帰納的および発想的なデータ処理や知識処理を含めた知識創造支援型データベースシステムの研究が重要になると考える。

２．研究の目的

知識創造支援型データベースシステムの研究を行うために、知識創造支援機能の構成法、知識創造支援処理の効率化法、ソーシャルネットワークを対象として知識創造を支援する方法、知識創造支援機能を応用する方法について、明らかにする。

３．研究の方法

記憶・演繹・帰納・発想やコミュニケーションの成立に重点をおいた知識創造支援型データベースシステムの研究を行うためには、具体的な例が必要であり、その例としてバイオインフォマティクス、テキスト処理応用、Web アプリケーションなどをとりあげ、(1)知識創造支援型データベースシステムの構成法の研究、(2)知識創造支援型データベース処理の効率化の研究、(3)知識創造過程におけるコミュニケーション成立を支援する方法の研究、(4)知識創造支援型データベースを応用する方法の研究を行う。

４．研究成果

研究の方法に掲げた４点について、以下のような成果が得られた。

（１）知識創造支援機能の構成法

この構成法を明らかにするために，汎化パターンの抽出法、類似構造検索法、の２つの観点から研究を行った。

① 汎化パターンの抽出法

類似部分文字列検索の結果として得られるミスマッチクラスタが多くの類似文字列を含むことに注目し、それを説明する正規表現を導出する方法[雑誌論文⑤、学会発表⑬]を初めて明らかにした。その導出方法に関して、反復精密化法と段階的一般化法と考案し、それらの特徴を明らかにした。図１は、ミスマッチクラスタとして、{ (1, <ABF>), (2, <AEC>), (3, <AEF>), (4, <DBF>), (5,<DEC>), (6, <DEF> }が与えられているとし、段階的一般化法により正規表現を探索する例である。図の二重丸が解答 {<[AD][BE]F>, <[AD]E[CF]>}を意味する。

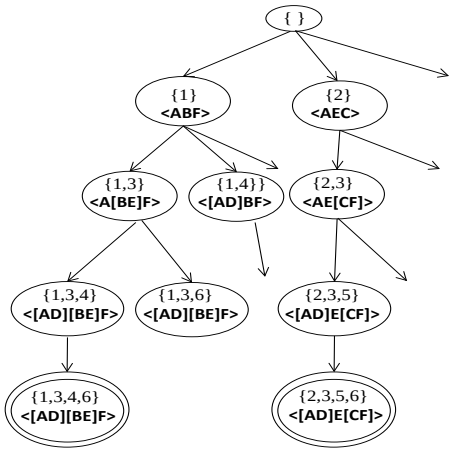

図１　正規表現の導出例

ミスマッチクラスタは、配列データベースに対してギブスサンプリングを適用することによっても得られることから、KDD オブジェクトの１つであり、KDD オブジェクトからさらに規則性を見出す研究は、帰納的データベースの研究上、大変意義深いといえる。その他、予め、ミスマッチクラスタにマルチプルアラインメントを適用しておけば、正規表現の計算精度の向上につながることについても明らかにした[学会発表⑨]。

汎化パターンの抽出法についての今後の展開としては、ギブスサンプリングの精度向上という観点から汎化パターンを有効利用する方法の研究や汎化パターンの冗長性を除去する研究などが挙げられる。

② 類似構造検索法

空間的な座標配列データに対する構造アラインメントを計算する問題に取り組み，その問題を解くことに成功した[雑誌論文②、学会発表③]。空間データの構造アラインメントは、空間的な類似構造をみつけだすのに有力な手段であるが、改良版 EO で高精度に解く方法を明らかにした。今後の展開としては、類似部分構造の自動抽出問題[学会発表⑪]の取組みを活かし、島モデルの導入などによる効果的な最適化手法や分散並列処理による

高速化の研究が残されている。

（２）知識創造支援処理の効率化法

この効率化法を明らかにするために、並列処理、索引構造の２つの観点から研究を行った。

① 並列処理

ミスマッチクラスタから高速に汎化パターンを抽出する段階的一般化法に着目した。この段階的一般化法の並列化に当たっては、マルチコア PC クラスタを使用し、その中で階層的タスクプールを考案した。その結果、良好な並列性能を引き出すことに成功した[雑誌論文④]。図２は、Kringle データセットを用いて、並列性能を測定した結果である。他のデータセットでも同様の結果が得られており、階層的キャッシュベースのランダムタスク・ステイル法と呼ばれる負荷分散法を使用した場合が負荷分散法を利用しない方法に比べて高い並列性能がえられた。

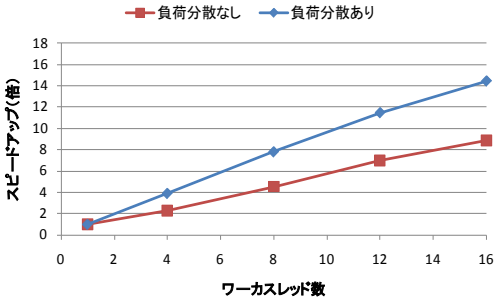

図２　並列化による速度向上比

② 索引構造

大規模な文字列データや座標配列データをとりあげ、サフィックス木の概念に基づいて索引構造を構築する方法について検討をおこなった[学会発表①⑤⑩⑫]が、今後、細部にわたる評価が一部残っている。

（３）知識創造過程におけるコミュニケーション成立の支援法

さまざまな状況におけるコミュニケーション成立をめざし、グラフィカルに異種木構造データベースの調停問題を効率的に解く方法やソーシャルネットワークを活用する方法に取り組んだ。

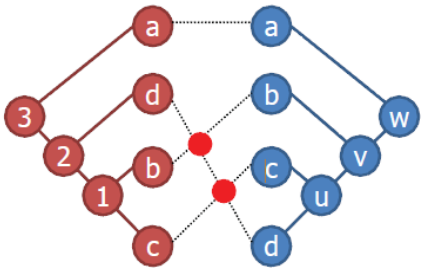

図３　調停グラフの交差数減少問題

① 調停問題をグラフィカルに解く方法

図３の例に示すような調停グラフの交差数減少問題に着目した。調停グラフは、異種木構造データどうしの違いをグラフィカルに把握するのに有用であり、交差数を減らす最適化問題を解くことが重要になっている。この問題に取り組んだ結果、島モデルを用いて EO を改良することによりこの最適化問題を効果的に解くことに成功した[雑誌論文①、雑誌論文⑦、学会発表②⑧]。図４は、DMEO 法（島モデルを用いて EO を改良した方法）と MGG 法の性能比較をした結果である。DMEO 法は MGG に比べて、良い性能を持つことがわかった。なお、評価実験では Moss データセットを用いた。

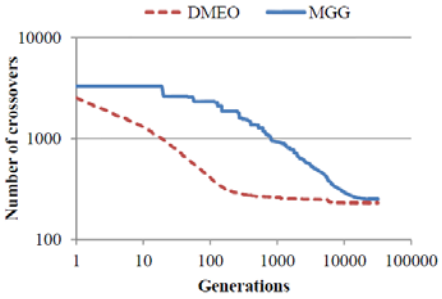

図４　世代に対する交差数の関係

② ソーシャルネットワークを活用する方法

知識情報システムとの連携に有用な Twitter や Blog に注目して、バースト性検出、コミュニティ抽出に関わる問題に取り組んだ。その結果、バースト性検出については、画像付きツイートからトピックスを抽出する方法[学会発表⑥]を初めとして、利用者の位置を考慮したバースト性検出アルゴリズム[学会発表⑦]を明らかにした。これを踏まえ、大規模文書ストリームからバースト性を高速検出するための並列化に成功した[雑誌論文③、学会発表④]。また、Blog 空間からコミュニティを自動抽出する研究を行い、重なりを考慮したコミュニティ抽出の方法を明らかにした[雑誌論文⑥⑧]。

今後の展開としては、ソーシャルネットワークを利用したコミュニケーション成立をめざし、知識創造支援システムとソーシャルネットワークとを上手く連携させる方法の研究があげられる。

（４）知識創造支援機能を応用する方法

産学連携によるマッチングセッションや広島市役所での展示会などでのポスター展示などを通じ、図書推薦、Web マイニング、バイオインフォマティクス、テキスト処理などの分野への先進的な応用可能性について

検討を行った。さらに、これまで得られた知見の一部を教科書としてまとめ[図書①②③]、研究者や学生など多くの人々が、その応用可能性について考える機会を作った。今後は、引き続き、知識創造支援機能の応用に関する研究を進めていきたい。

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８件）

① Keiichi Tamura, Hajime Kitakami, and Akihiro Nakada: Distributed Modified Extremal Optimization for Reducing Crossovers in Reconciliation Graph, The special issue of International Association for Engineers(IAENG)、Engineering Letters、査読有、Volume 21、Issue 2、 June 2013、 pp.81-88、http://www.engineeringletters.com/issues_v21/issue_2/EL_21_2_05.pdf.

② 中田 章宏、田村 慶一、北上 始、高橋 誉文: CMO問題に対する改良版EOを用いた発見的解法、情報処理学会論文誌数理モデル化と応用(TOM)、査読有、Vol.6、 No.3、2013年、 12 pages.

③ Kaishi Hirahara、Keiichi Tamura、Hajime Kitakami、and Shingo Tamura: Parallel Processing of Burst Detection in Large-Scale Document Streams and Its Performance Evaluation、GSTF International Journal on Computing、The Global Science and Technology Forum (GSTF) publishes、査読有、Vol.2、 No.4、 March 2013、 7 pages.
DOI: 10.5176/2251-3043_2.4.206

④ Yagi Shinpei、Keiichi Tamura、and Hajime Kitakami: Parallel Processing for Stepwise Generalization Method on Multi-Core PC Cluster、Special Issue on ″Advanced Soft Computing Methodologies and Applications in Web Intelligences、″International Journal of Knowledge and Web Intelligence (IJKWI)、Inderscience Publishers、査読有、Vol. 3、 No. 2、2012、 pp.88-109.
DOI: 10.1504/IJKWI.2012.050282

⑤ 田村 慶一、木村 浩明、荒木 康太郎、北上 始:段階的一般化法によるミスマッチクラスタを表現する最小汎化集合の効率的抽出、電子情報通信学会和文論文誌Ｄ「データ工学特集号」、電子情報通信学会、査読有、Vol.J93-D、No.3、2010年3月、pp.189-202、http://harp.lib.hiroshima-u.ac.jp/handle/harp/7007.

⑥ 高木 允、田村 慶一、森 康真、北上 始:密な部分構造抽出のための階層的凝集型クラスタリング手法、日本データベース学会論文誌、査読有、Vol.7 No.1、2008年6月、 pp.275-280.

⑦ 田村 慶一、森 康真、北上 始: Extremal Optimizationによる調停グラフの交差数減少、情報処理学会論文誌数理モデルと応用(TOM)、査読有、Vol.4、 No. SIG4 (TOM20)、 2008年3月、pp. 105-116.

⑧ 高木 允、森 康真、田村 慶一、北上 始:ブログユーザ空間からの重複を許した頻出コミュニティ抽出法、情報処理学会論文誌数理モデルと応用(TOM)、査読有、Vol. 49、 No. SIG4(TOM20)、2008年3月、pp.93-104.

〔学会発表〕（計13件）

① Yosuke Watanuki、Keiichi Tamura、Hajime Kitakami、Yoshifumi Takahashi: Parallel Processing of Approximate Sequence Matching using Disk-based Suffix Tree on Multi-core CPU、The 6th International Workshop on Computational Intelligence & Applications 2013(IWCIA2013)、査読有、July 2013、6 pages.

② Keiichi Tamura、Hajime Kitakami、and Akihiko Nakada: Distributed Modified Extremal Optimization for Reducing Crossovers in Reconciliation Graph【Certificate of Merit for The 2013 IAENG International Conference on Artificial Intelligence and Applications】、Proceedings of the International Multi-Conference of Engineers and Computer Scientists 2013、The 2013 IAENG International Conference on Artificial Intelligence and Applications (ICAIA)、 査読有、13-15 March 2013、pp.1-6、 The Royal Garden Hotel、 Kowloon、 Hong Kong.

③ Akihiko Nakada、Keiichi Tamura、and Hajime Kitakami: Optimal Protein Structure alignment using Modified Extremal Optimization、 The 2012 IEEE International Conference on Systems、Man、and Cybernetics (IEEE SMC 2012)、査読有、 14-17 October 2012、Seoul in Korea、 pp.697-702.

④ Kaishi Hirahara【Best Student Paper Award】、Keiichi Tamura、Hajime Kitakami、and Shingo Tamura: Parallel Processing of Burst Detection in Large-Scale Document Streams、Proceedings of 3rd

Annual International Conference on Advances in Distributed and Parallel Computing、September 17-18 in 2012、pp.60-65、Bali in Indonesia.

⑤ グエン フーン バック、高橋 誉文、田村慶一、北上 始：大規模文字列データベースに対する索引方式の考察、査読無、IEEE SMC Hiroshma Chapter若手研究会、 2012年7月14日、 pp.14-16.

⑥ Shingo Tamura、Keiichi Tamura、Hajime Kitakami 、 and Kaishi Hirahara: Clustering-based Burst-detection Algorithm for Web-image Document Stream on Social Media、 The 2012 IEEE International Conference on Systems、Man、and Cybernetics (IEEE SMC 2012)、査読有、 14-17 October 2012、Seoul in Korea、 pp.60-65.

⑦ Keiichi Tamura and Hajime Kitakami: Location-Based Burst Detection Algorithm in Spatiotemporal Document Stream 、 The 2012 International Conference on Data Mining (DMIN12) 、CSREA Press、査読有、July 16-19、2012、pp.195-201、Las Vegas in USA.

⑧ Natsumi Hara 、 Keiichi Tamura 、and Hajime Kitakami: Modified EO-based Evolutionary Algorithm for Reducing Crossovers of Reconciliation Graph、Proceedings of the Second World Congress on Nature and Biologically Inspired Computing (NaBIC2010)、 IEEE Computer Society Press 、 査 読 有 、December 15-17、 2010、pp.169-176.

⑨ Kazuya Miyahara、 Hajime Kitakami、Yoshifumi Takahashi、 Keiichi Tamura and Susumu Kuroki: Mining Minimum Generalized Set Based on Multiple Alignments from Mismatch Cluster 、Proceedings of The 2010 International Conference on Bioinformatics and Computational Biology (BIOCOMP'10)、CSREA Press、査読有、 July 2010、 Vol. Ⅰ、 pp.35-41、Las Vegas in USA.

⑩ Yoshifumi Takahashi、 Susumu Kuroki、and Hajime Kitakami: Efficient Query Processing in Protein Structure Databases、 Proceedings of the 2nd International Workshop with Mentors on Databases 、 Web and Information Management for Young Researchers (iDB Workshop 2010)、 査読有、 2nd - 4th August、 2010、pp.47-55.

⑪ 河野 修久、田村 慶一、森 康真、北上 始: ギブスサンプリングとアラインメント処理に基づく類似部分配列の抽出方式、情報処理学会・数理モデル化と問題解決（MPS）およびバイオ情報学（BIO)、査読無 、 Vol.2009-MPS-76 No.46 および Vol.2009-BIO-19 No.46、 2009年12月、8 pages.

⑫ Yusuke Sawada、Keiichi Tamura、Kotaro Araki 、 Makoto Takaki 、 and Hajime Kitakami: Parallel Construction Method of a Disk-Based Suffix Tree on a PC Cluster、 Proceedings of the 2008 International Conference on Parallel & Distributed Processing Techniques & Applications (PDPTA'08) 、CSREA Press、査読有、 July 14-17 in 2008、Vol.II、pp.797-803、 Las Vegas in USA.

⑬ Hiroaki Kimura 、 Hajime Kitakami 、Kotaro Araki、 and Keiichi Tamura: A Stepwise Generalization Method for Extracting Minimum Generalized Set from Mismatch Cluster、 Proceedings of the 2008 International Conference on Bioinformatics and Computational Biology (BIOCOMP'08) 、 CSREA Press、査読有、Vol.II、 July 14-17 in 2008、pp.998-1004、 Las Vegas in USA.

〔図書〕（計3件）

① 北上 始：第6章 データベース、情報とネットワーク社会、オーム社、2011年12月、pp.119-148.

② 北上 始、黒木 進、田村 慶一：データベースと知識発見、コロナ社、2013 年 10月、176ページ.

③ 森 康真：第9章 問題解決と情報の管理、一般教育の情報（北上 始 編)、あいり出版、2013年10月、pp.145-159.

６．研究組織

(1)研究代表者

北上　始（KITAKAMI HAJIME）

広島市立大学・情報科学研究科・教授

研究者番号：50234240

(2)研究分担者

田村　慶一（TAMURA KEIICHI）

広島市立大学・情報科学研究科・准教授

研究者番号：80347616

黒木　進（KUROKI SUSUMU）

広島市立大学・情報科学研究科・准教授

研究者番号：20225288

（H24：連携研究者）

森　康真（MORI YASUMA）

広島市立大学・情報科学研究科・助教

研究者番号：60264959

（H24：連携研究者）