戦略的創造研究推進事業
発展研究（SORST）

研究終了報告書

研究課題
「パターン照合とテキスト圧縮に基づく
高速知識発見技術に関する研究」

研究期間：平成１５年１０月　１日～
平成１９年　３月３１日

竹田　正幸
（九州大学、教授）

１．研究課題名

パターン照合とテキスト圧縮に基づく高速知識発見技術に関する研究

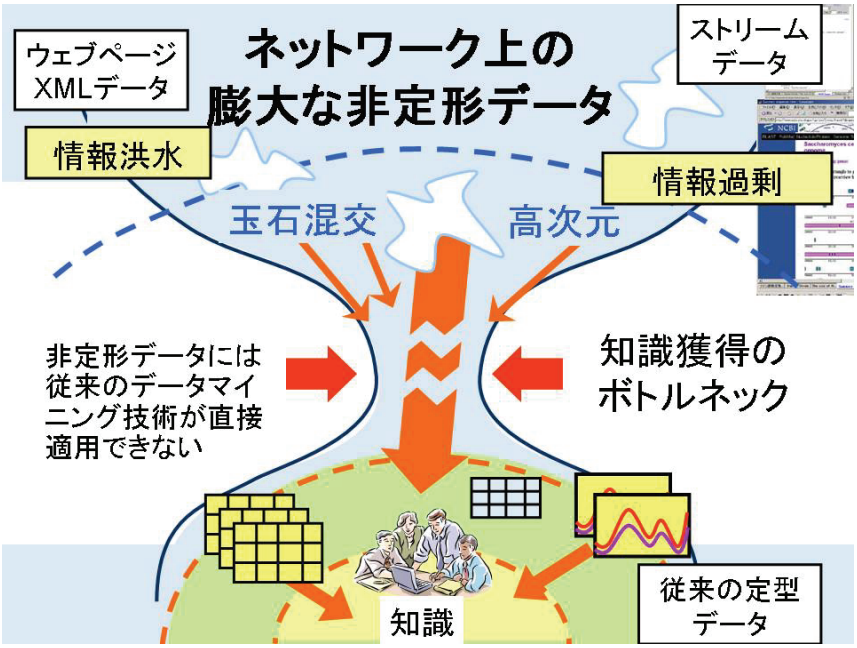
２．研究実施の概要

背景

情報洪水時代の到来：計算機機器の低価格化やネットワーク技術の進展を背景として，様々な情報の機械可読化が進み，情報洪水時代が到来している．物理学・化学・生物学等の自然科学分野における実験・観測データは巨大化してきており，計算機なしでは解析が不可能である．また，インターネット上に溢れる Web ページも，全体でひとつの巨大データベースであるかのような様相を呈している．さらに，企業内の販売・顧客データや機械可読文書を蓄積した XML アーカイブ等も，巨大化の一途を辿っている．

非定型データの扱い：ここで問題となるのは，これらの膨大なデータの多くが，従来の関係型データベースのような定型的データではなく，**定まった形式を持たないテキストデータ**である点である．従来の関係型データベースが“hard to publish, easy to query”（公開は困難，質問は容易）であるのに対し， テキストデータは，“easy to publish, hard to query”と言われる．すなわち，“easy to publish”であるテキストデータは，関係型データベースでは想像もつかなかった速度で増加し続けるのである．

眠れる知識の発掘：これらの巨大なテキストデータは，知識の宝庫である．しかし，従来のデータベース技術は専ら定型的なデータを対象として発達したため，テキストデータのような非定型データを扱うための基礎技術が十分に確立されているとは言い難い．そこで，**データ中に潜む宝を掘り起こし，有効に活用するための新しいデータアクセス技術の確立**が待望されている．



本研究の特色

第１の特色：テキストは，そのままでは，何の構造も持たない文字の連鎖である．テキストデータマイニング研究の多くは，この文字の連鎖を従来の定型データの枠に当てはめようするが，文字の連鎖にはそれに相応しい処理方式をとるべきである．本研究では，陽には構造を持たないテキスト中に潜む規則性を捉えるモデルとして，様々な形式のパターンを考え，**パターン照合技術を核とした知識発見システムの構築を目指す．この点が本研究の第１の特色である**．

第２の特色：ここで知識発見とは，データ中に潜む規則性を計算機によって見いだすことを言う．一方，情報科学の古典的研究分野のひとつであるデータ圧縮も，データ中の持つ規則性に基づいてその記述長を抑える技術である．Rissanen の最小記述長原理に基づいた機械発見手法がしばしば良い成果を上げていることからも判るとおり，データ圧縮と機械発見の二つには深い関係がある．従来データ圧縮の研究は，圧縮率の向上と圧縮・展開に要する計算時間の短縮を主な目標としてきた．そこで，これとは別の視点，すなわち，機械発見の視点からデータ圧縮技術を眺め直すことで，これまで看過されていた技術が脚光を浴びる可能性がある．すなわち，**データ圧縮技術を機械発見の観点から再評価し，その知見を積極的に援用して知識発見基盤技術の確立を目指す点が，第２の特色である**．

第３の特色：知識発見システムの成功のカギは，**人間によるシステムへの介入**であると言われる．データマイニング研究では，入力データに対して仮説を出力するアルゴリズムにばかり目がいきがちであるが，出力された仮説に専門家が意味付けしてこそ，有益な知識の発見が可能となる．「さきがけ研究 21」の経験では，専門家の反応は，設定した仮説空間の質に左右される．すなわち，仮説空間内の個々の仮説の表現力が豊かで，対象とする問題領域の性質をうまく反映していなければ，専門家は出力された仮説に対して積極的にコミットし得ない．一方，計算時間の面からは，仮説は単純であるほどよい．この相反する二つの要件のあいだでうまくバランスをとりながら仮説空間を設定することが必要である．そこで，**仮説空間設定への指針を与えるために，様々な仮説空間に対するパターン発見問題の計算量の階層を究明する．この点が本研究の第３の特色である．**

研究成果

以下に示す３つの項目ごとに研究を行い，優れた研究成果をあげた．

I．非定型データの高速パターン照合技術

代表者の有する系列データを扱う高速パターン照合技術を核に据えることにより，非定型データへの有効なアクセスメソッドとして，軽量かつ高速な XML ストリーム処理器 XAXEN を開発した．XAXEN は数千～数万のクエリを同時に処理することが可能で，ストリーム処理手法として有名な XMLTK, YFilter と比較して，**実行速度で約 4～6 倍，メモリ使用量で約 6 倍以上**の圧倒的な性能を達成した．また，市販の XML データベース管理ソフトウェアであり，国内シェアの１，２を争う Tamino および NeoCoreXMS との比較を行った結果，使用するメモリ・ディスク容量，処理速度やその安定性，およびクエリ数に関する頑健性において，**本方式が圧倒的に優れている**ことが判明した．

II．データ圧縮に基づく高速非定型データ処理技術

非定型データのパターン照合処理を高速化するためのもうひとつの技術として，代表者らが世界に先駆けて開発した「テキスト圧縮による高速化」技術がある．さらなる高速化を図るため，(a)新たな圧縮パターン照合方式の開発と，(b)それに適した新たな圧縮法の開発を目指した．(a)については，コラージュシステム上のパターン照合の枠組みを拡張し，実用的観点から優れた変数系列の符号化方式を開発し，それに合わせたパターン照合アルゴリズムを設計した．これにより，たとえば英文テキストの場合，**従来手法では 60%程度にしか実行時間を短縮できなかったものを 40%程度にまで短縮することに成功**した．

(b)については，コラージュシステムの重要な部分族である正規コラージュシステムに対して文法サイズ最小化問題を考え，これが NP 困難であることから，この問題に対する現実的な解として長さ優先置換法に基づく圧縮スキーマに着目した．この圧縮は，ナイーブ

な方法では $O(n^4)$時間を要し，Minimal Augmented Suffix Tree とよばれる凝ったデータ構造を用いても $O(n^2 \log n)$時間を必要とするものである．本研究では，**文字列の組合せ的性質に関する知見を駆使することでこれを劇的に削減**し，$O(n)$時間・領域で計算するアルゴリズムを開発した．

III．非定型データからのパターン発見技術

理論的側面：最適弁別パターン発見問題について，新たなパターン族を導入し，それぞれについて効率的なパターン発見アルゴリズムを開発した．また，そのアルゴリズムを高速化するためのデータ構造の開発を行った．

応用的側面：応用としては，音楽データ・言語データ・薬学データ・税関の申告データなどさまざまな分野のデータに適用した．特に，医薬品情報学分野の研究者との連携により，医薬品商標名の類似性を定量化する類似性指標および類似度算出方式を開発し，**厚生労働省研究チームによる既存の指標を大きく上回る性能**を得た．また，文字列の「異質性」という概念を導入してそれを定量化し，**社会的ニーズの極めて大きいスパム自動抽出へ応用して高い精**度を得た．

３．研究構想

研究目的

本研究では，知識発見処理に必要な要素技術の確立を目指す．その際，計算量に徹底した配慮をしながら研究を遂行する．

知識発見処理の高速化の効用は，単なるスピードアップにとどまらない．すなわち，あまりにも時間がかかりすぎるため**「事実上の不可能」として退けていた大がかりな処理が，高速化によって初めて現実のものとなる**ことがある．また，もう一つの効用として，**高速化により，知識発見処理が，パソコン（PC）などの比較的安価な計算機によって実行可能**になることもある．実際，「さきがけ研究 21」において，当初は大規模で高価な計算機システムでなければ実行できなかった処理を，安価な PC でも実行できるところまで改善できた．

本研究で開発する**非定型なテキストデータからの知識発見システムは，「情報洪水時代」を生き抜くためのサバイバルキット**たりうる．このサバイバルキットは，大規模な計算機システムを有する一部の特権的ユーザのみが利用できるものであってはならない．計算量に徹底した配慮をした研究を遂行することにより，処理の高速化と省メモリ化が達成されれば，このサバイバルキットを，誰もが気軽に利用できるようになることも夢ではない．

研究の方法

高速化といっても，無論，誰も使わないような処理を高速化しても始まらない．そこで，本研究では，「さきがけ研究 21」の３年間と同様，理論と実用の両面から研究を進める．すなわち， 以下の２つを繰り返すのである．

- 理論的研究で得た技術を現実の問題に適用する．
- 実用的研究の場面で得た知見を，理論研究にフィードバックし，必要な基盤技術を開発する．

この研究スタイルにより，理論・実用のいずれにも偏らない有益な研究が実現する．具体的には，次の３つを研究項目として研究を進める．

I　非定型データの高速パターン照合処理技術．
II　データ圧縮に基づく高速非定型データ処理技術．
III 非定型データからのパターン発見技術．

ここで，鍵となるのが**データ圧縮技術**である．II では高速化のための手段として圧縮技術を用い，III では，「発見＝圧縮」という観点から，圧縮としての発見技術を追求する．

知識発見処理は，理論的に計算困難であることが多いが，それでも実用的な計算方式を開発しなければならない．これには，従来のように，高性能の CPU と大容量の記憶装置を

備えた巨大かつ高価な計算機システムを用いるのではなく，安価な PC 群を用いた分散計算方式が有効と考えられる．

そこで，本研究においては，設備備品費の大部分をブレードサーバの購入に充てる．このブレードサーバは，従来のラックマウント型のサーバと比べて集積度が高く，省スペースのみならず省電力の効果が大きいことから，近年注目を集めている．研究の進展に伴い，随時，サーバブレードを追加し，より大規模な分散環境を構築し，計算機実験を行う．

４．研究実施内容

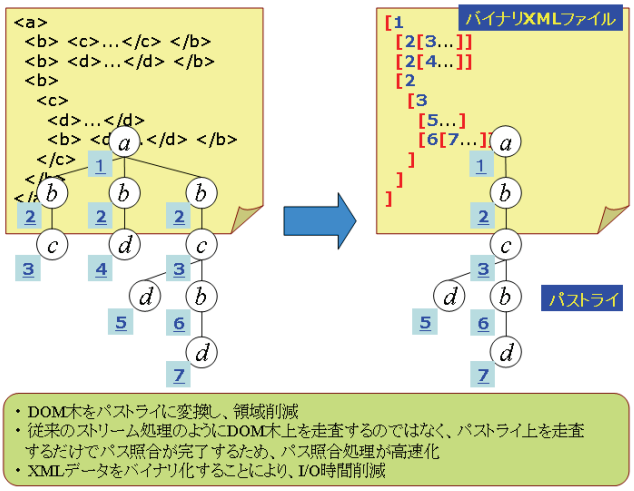
(1)実施の内容

本研究は，情報爆発時代において出現した，「大量・非定型・不均質」という特徴をもつ新しいタイプのデータベースに対する基盤技術の開発を目的とし，以下の３つの項目について研究を行なった．

Ⅰ．非定型データに対する高速パターン照合技術

(A) 索引を用いる常識的な従来型処理方式と，代表者らによる高速一方向逐次処理技術との比較を詳細に行った．その結果，100MB 程度のデータを対象とした場合，クエリ数がある程度以上になると，逐次処理方式がむしろ高速であるという，**常識を覆す事実**を示した．

(B)上述の高速一方向逐次処理技術を基礎とする超高速系列データ走査手法と，パストライを活用した構造パターン照合手法を結合することによって，高速かつ軽量な XML ストリーム処理器 XAXEN を開発した．XAXEN は数千～数万のクエリを同時に処理することが可能で，ストリーム処理手法として有名な XMLTK, YFilter と比較して，**実行速度で約 4～6 倍，メモリ使用量で約 6 倍以上**の圧倒的な性能を示す．



さらに，XMark の提供するベンチマークを用いて，市販の XML データベース管理ソフトウェアであり，国内シェアの 1,2 を争う Tamino および NeoCoreXMS との比較を行った．その結果，使用するメモリ・ディスク容量，処理速度やその安定性，およびクエリ数に関する頑健性に関して，**本方式がこれら商用の DB システムに比べ圧倒的に優れていることが判明**した．

(C)上述の方式に基づいた並列分散検索システムを構築した．すなわち，各サーバブレードがそれぞれの担当すべきデータ断片について逐次処理を行うものである．この方式は，大規模なデータを扱う場合であっても，クエリ数が一定以上になると，索引構造等を用いる従来型処理方式と比べて高速になる．

(D) HTML 文書や XML 文書などの順序木の蓄積から構造に関するパターンを取り出すための基礎技術として，正則生垣とよばれるパターン木族に対する効率的な木パターン照

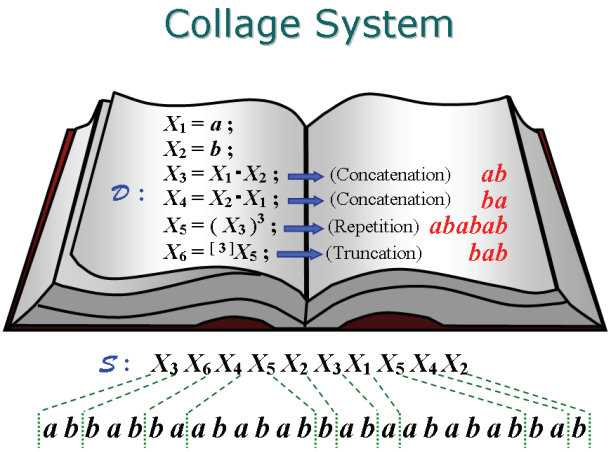
合アルゴリズムを開発した[27].

(E)接尾辞木は文字列照合においてカギとなるデータ構造であり，バイオインフォマティクスやデータ圧縮など幅広い応用分野で用いられている．疎接尾辞木は接尾辞木の変種であって，入力文字列の接尾辞集合の部分集合のみを表現する．本研究では語接尾辞木を扱った．語接尾辞木は，疎接尾辞木のひとつである．D を語の辞書とし，w を D^+に属する文字列とする．すなわち，w は D に属する語 $w_1 \ldots w_k$ から成る系列である．D に関する語接尾辞木とは，$w_i \ldots w_k$ という形をした k 個の接尾辞だけを表現する経路圧縮トライをいう．典型的な応用例としては，自然言語文書に対する語あるいは句レベルでの検索があげられる．Andersson らは語接尾辞木を構築するアルゴリズムを示した．このアルゴリズムは，平均時 $O(n)$時間で $O(k)$領域を用いて動作する．しかし，最悪時においても $O(n)$時間・$O(k)$領域で語接尾辞木を構築する問題は，10 余年もの間未解決問題であった．

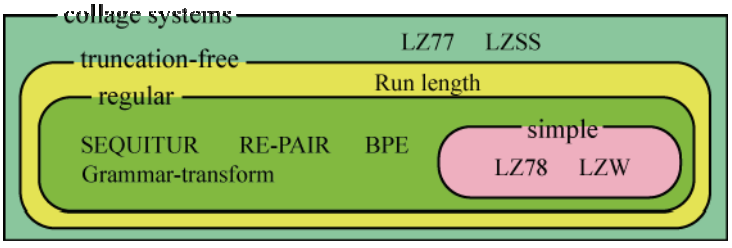
本研究では，**この未解決問題を解く新しい語接尾辞木構築アルゴリズムを開発**した[8]．本アルゴリズムは，最悪時においても $O(n)$時間・$O(k)$領域で動作する．また，このアルゴリズムはオンライン，すなわち，入力文字列の文字を左から右に 1 文字ずつ逐次処理するという好ましい特徴をも有している．また，同様のアイデアが DAWG に適用可能であることを示し[5]，さらには，二つを結合して CDAWG に適用し，よりコンパクトな索引構造を得た[7].

II.　データ圧縮に基づく高速非定型データ処理技術

(A)「さきがけ 21」の研究では，**コラージュシステム**と名づけた文字列表現体系を提案し，既存の圧縮法を統一的に扱うことを可能とし，汎用の圧縮パターンアルゴリズムを開発した．コラージュシステムとは，右に示すように，辞書構造 D とトークン列 S の対である．本研究では，このコラージュシステムのもとで，**文字列処理高速化と省ストレージ化の両方の観点から性能のよい，新しい圧縮方式を開発する**とともに，それに対応した**新たな圧縮パターン照合方式の確立**を目指した．コラージュシステム上のパターン照合の枠組みを拡張し，実用的観点から優れた変数系列の符号化方式を開発し，それに合わせたパターン照合アルゴリズムを設計した．これにより，たとえば英文テキストの場合，**従来手法では 60%程度にしか実行時間を短縮できなかったものを 40%程度にまで短縮**することに成功した．



(B) コラージュシステムの重要な部分族である正規コラージュシステムに対して文法サイズ最小化問題を考え，これが NP 困難であることから，この問題に対する現実的な解として長さ優先置換法に基づく圧縮スキーマに着目した．この圧縮は，ナイーブな方法では $O(n^4)$時間を要し，Minimal Augmented Suffix Tree とよばれるかなり凝ったデータ構造を用いても $O(n^2 \log n)$時間を必要とするものである．**本研究では文字列の組合せ的性質に関する知見を駆使することによりこれを劇的に削減し，$O(n)$時間・領域で計算するアルゴリズムの開発に成功**した[30,3].



(C) コラージュシステムの部分族である単純コラージュシステムを対象に，完全圧縮パターン照合問題を効率的に解くアルゴリズムを開発した[20,18].

III. 非定型データからのパターン発見技術

理論的側面

(A) 最適パターン発見問題について，ウインドウ幅制限や近似照合などに基づいた新たなパターン族を導入し，効率的なパターン発見アルゴリズムを開発するとともに，アルゴリズム高速化のためのデータ構造の開発を行った[28,6]．二つの部分文字列パターン p, q のすべてのブール結合（p∧¬q など）という新たなパターン族を対象とし，最適パターン発見問題を解く効率的アルゴリズムを開発した[19]. このアルゴリズムは最適パターン発見の一般化である相関パターン発見問題へ適用することが可能である．式 p∧q, p∧¬q における部分文字列パターン p, q それぞれの生起の距離に制限を導入した問題にも取り組み，効率的アルゴリズムを与えた[22]．「局所関連性」という新しい概念を導入し，それに基づいたいくつかのパターン族に対する効率的パターン発見手法を与えた[4].

(B) 最適弁別パターン発見問題を一般化し，文字列とそれに関連付けられた数値属性値との対の集合から最適パターンを発見する問題に取り組んだ[13]．パターンの「良さ」として，個々の文字列におけるパターンの生起回数とその文字列に関連付けられた数値属性値との間の相関を用いる．部分文字列族を仮説空間とした場合に対して，接尾辞木に基づく二つの効率的アルゴリズムの開発に成功した．またより複雑なパターン族を扱う際に用いることのできる一般的な分枝限定手法を示した．

(C) VLDC パターン族に対するパターン発見について，60 台のブレードを備えたブレードサーバを用いた並列分散計算によるパターン発見の高速化を達成した．さらに，開発した枝刈手法の効果を詳細に調べ，アルファベットサイズが増大しても実用時間で処理を完了できることが判明した．これにより，従来は事実上適用できなかった漢字仮名混じりの日本語テキストからのパターン発見の可能性を示した．

応用的側面

(D) MIDI データからメロディ，リズム，コード進行のデータを抜き出し，これに上述のパターン発見手法を適用することにより，作曲家の特徴をあらわすパターンを抽出することに成功した．また，J-POP と演歌を比較することにより，両者の差異に対応するパターンを抽出した．

(E) 医薬品情報学分野の研究者との連携により，医薬品商標名の類似性を定量化する類似性指標および類似度算出方式を開発した．薬剤師らによる投薬ミスのデータを用いた検証では，厚生労働省研究チームによる既存の指標を大きく上回る性能を得た．この結果は，情報科学分野の成果であるばかりでなく，薬学分野第一級の学術雑誌である「薬学雑誌」に掲載された[10].

(F) 門司税関の協力を得て，通関データからの不正輸出等の検出を目的としたデータマイニングの問題に取り組み，現在人手に頼るしかない輸入業者の特徴抽出への足がかりを得た．

(G) 与えら得た文字列集合における部分文字列について，「異質性」という概念を導入し，Blumer らによる同値関係に基づいて異質性の定量化を行なった．また，その値の計算を高速かつ省メモリで行なうためのデータ構造とアルゴリズムを開発した[1]．さらに，その応用として，Web スパム検出の問題に適用し，単独の手法としては高い精度を得た[9].

(2)得られた研究成果の状況及び今後期待される効果

研究成果の状況：以上述べてきたように，本研究では知識発見基盤技術の開発について，理論と実用の両面から研究を行なっており，いずれについても満足のいく結果を得ることができた．本研究では，「非定型データ」を基本的に単なる文字の連鎖として扱っており，データの内容は問わない．したがって，Web ページなどの自然言語文はもとより，遺伝子情報，音楽情報などにも適用できるなど，応用範囲は極めて広い．

今後期待される効果： 研究項目 I で開発した高速化は劇的であり，かつ使用するメモリ

を小さく抑えられる点が特徴的である．したがって，この技術により，計算資源(CPU パワーやメモリ・ディスク容量)の乏しい組込み機器であっても情報獲得・知識発見処理を行なうことが可能となる．来るべきユビキタス社会においては，現在の携帯電話よりもさらに小型の情報機器が多量に普及することとなり，現在よりも大量の情報をすばやく処理することが重要となる．したがって，この技術は，ユビキタス情報機器の組み込みアプリケーションにとって核となる技術へとつながるものと期待できる．

研究項目 II の圧縮パターン照合技術は，I の高速化技術と結合することにより，さらなる高速化へ寄与するものである．情報爆発時代においては，データ格納およびデータ転送のコストの両方を抑えることが必須となり，データ圧縮の重要性をますます顕在化する．通常ならば，圧縮のデメリットとしてデータを伸張するためのコストがかかるが，圧縮したデータを伸張することなく処理する本方式は，この問題を根本から解決する画期的な技術である．

研究項目 III の知識発見の基盤となる技術により，非定型データを対象とした，より知的な情報アクセスが可能となる．I, II, III の研究成果をさらに発展させることにより，「新世代型データベース基盤技術」を構築する計画である．

５．類似研究の国内外の研究動向・状況と本研究課題の位置づけ

研究項目 I で扱った高速パターン照合技術については，世界の各所で盛んに研究され，その成果は CPM (Combinatorial Pattern Matching)や SPIRE (String Processing and Information Retrieval)などの当該分野トップレベルの国際会議で発表されている．代表者はそのいずれにも多数の論文を発表しており，また，プログラム委員長やプログラム委員を務めるなどその運営にも係っており，代表者らの研究グループは世界の研究拠点のひとつを形成していると目されている．

XML ストリーム高速処理に関しては，世界各所のデータベース研究者およびデジタル文書研究者らによって盛んに研究されている．代表者は，当該分野の研究者ではないが，上述の高速パターン照合技術を駆使して高速 XML ストリーム処理器 XAXEN を開発し，既存技術を大きく上回る性能をもつことを示した．代表者らのようなアプローチは，当該分野においてはかなり独特なものであり，ほとんど類をみない．

研究項目 II で扱った圧縮パターン照合技術は，ハイファ大学(イスラエル)，チリ大学などでも盛んに研究されているが，代表者は当該分野の第一人者と目されており，著名な Springer 社から刊行予定の Encyclopedia of Algorithms(アルゴリズム百科事典)において圧縮パターン照合の項目を執筆している．今回，圧縮率と照合時間短縮率の両方を劇的に改善することができたが，この結果は，世界の他の研究を圧倒するものである．

研究項目 III で扱った知識発見については，世界中のデータマイニングや発見科学・機械学習の研究者らによって盛んに研究されている．代表者らは文字列データ・系列データに対象を絞っているが，同様の研究は主としてバイオインフォマティクス寄りの研究者に多く見られる．しかし，代表者のように対象を文字の連鎖として抽象的に扱いつつ同時に幅広い応用を目指すものは少ない．代表者は，発見科学に関する代表的な国際会議である DS (Discovery Science)に第 1 回の 1998 年当時から毎年 1 篇以上の論文が採択されているが，これは「さきがけ研究 21」およびそれに続く本研究課題による研究成果である．この貢献が評価されて 2003 年より毎年プログラム委員を務め 2007 年についてはプログラム委員長を務めている．

６．研究実施体制

(1)体制

研究代表者が単独で行なう.

(2)メンバー表

氏　名	所　属	役　職	研究項目	参加時期
脇田早苗	派遣先	研究補助員	音楽データ，薬学データの整理・機械処理	平成 16 年 10 月～ 平成 19 年 3 月
池内昌子	派遣先	研究補助員	言語データの整理・機械処理	平成 16 年 4 月～ 平成 17 年 10 月 平成 18 年 5～9 月 平成 19 年 2～3 月

７．研究期間中の主な活動

(1)ワークショップ・シンポジウム等

年月日	名称	場所	参加人数	概要
なし				

(2)招聘した研究者等

氏　名（所属、役職）	招聘の目的	滞在先	滞在期間
なし			

８．発展研究による主な研究成果

(1)論文発表（英文論文　２８件　　邦文論文　２件）

[1] Kazuyuki Narisawa, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda. Efficient Computation of Substring Equivalence Classes with Suffix Arrays. accepted for *the 18th Annual Symposium on Combinatorial Pattern Matching* (CPM07), 2007.

[2] ○Shuichi Mitarai, Akira Ishino, and Masayuki Takeda. Light-weight acceleration for streaming XML document filtering. In *Proc. The Third IEEE International Workshop on Databases for Next-Generation Researchers (SWOD'07)*, 2007.

[3] ○Ryosuke Nakamura, Hideo Bannai, Shunsuke Inenaga, and Masayuki Takeda.
Simple Linear-Time Off-Line Text Compression by Longest-First Substitution.
In *Proc. Data Compression Conference'07 (DCC'07)*, pp.123-132, IEEE Computer Society, March 2007.

[4] Yasuto Higa, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda.
A New Family of String Classifiers based on Local Relatedness.
In *Proc. 9th International Conference on Discovery Science (DS2006), LNAI 4265*, pp. 114-124, Springer-Verlag, October 2006.

[5] Shunsuke Inenaga and Masayuki Takeda.
Sparse Directed Acyclic Word Graphs.
In *Proc. 13th International Symposium on String Processing and Information Retrieval (SPIRE'06), LNCS 4209,* pp. 61-73, Springer-Verlag, October 2006.

[6] Yasuto Higa, Hideo Bannai, Shunsuke Inenaga, and Masayuki Takeda.
Reachability on Suffix Tree Graphs.
In *Proc. Prague Stringology Conference 2006*, pp. 212-225, Czech Technical University, August 2006.

[7] Shunsuke Inenaga and Masayuki Takeda.
Sparse Compact Directed Acyclic Word Graphs.
In *Proc. Prague Stringology Conference 2006*, pp. 195-211, Czech Technical University, August 2006.

[8] ○Shunsuke Inenaga and Masayuki Takeda.
On-line Linear-time Construction of Word Suffix Trees.
In *Proc. 17th Annual Symposium on Combinatorial Pattern Matching (CPM'06)*, LNCS 4009, pp. 60-71, Springer-Verlag, July 2006.

[9] Kazuyuki Narisawa, Yasuhiro Yamada, Daisuke Ikeda, and Masayuki Takeda.
Detecting Blog Spams using the Vocabulary Size of All Substrings in Their Copies.
In *Proc. 3rd Annual Workshop on Weblogging Ecosystem*, May 2006.

[10] ○大谷壽一, 竹田正幸, 今田結城, 澤田康文. 医薬品の取り違えミスを防止するための薬名類似度の定量的指標の構築.
薬学雑誌 126(5):349-356, 2006.

[11] 石野 明, 竹田 正幸.
パスプルーニングと決定性有限オートマトンを用いたストリーム指向の XQuery 処理,
日本データベース学会 Letters, Vol. 4, No. 4, pp.17-20, 2006.

[12] Yusuke Ishida, Shunsuke Inenaga, Ayumi Shinohara, and Masayuki Takeda.
Fully Incremental LCS Computation.
In *Proc. of the 15th International Symposium on Fundamentals of Computation Theory (FCT2005)*, Lecture Notes in Computer Science 3623, pp. 563-574, August 2005.

[13] Hideo Bannai, Kohei Hatano, Shunsuke Inenaga, and Masayuki Takeda.
Practical Algorithms for Pattern Based Linear Regression.
In *Proc. of the 8th International Conference on Discovery Science (DS 2005)*, Lecture Notes in Artificial Intelligence 3735, pp. 44-56, October 2005.

[14] ○Hisashi Tsuji, Akira Ishino, and Masayuki Takeda.
A Bit-Parallel Tree Matching Algorithm for Patterns with Horizontal VLDC's.
In *Proc. 12th International Conference on String Processing and Information Retrieval (SPIRE 2005)*, Lecture Notes in Computer Science 3772, pp. 388-398,

Novemver 2005.

[15] Shunsuke Inenaga, Hiromasa Hoshino, Ayumi Shinohara, Masayuki Takeda, Setsuo Arikawa, Giancarlo Mauri, and Giulio Pavesi.
On-Line Construction of Compact Directed Acyclic Word Graphs,
Discrete Applied Mathematics, 146(2):156-179, 2005.

[16] Satoru Miyamoto, Shunsuke Inenaga, Masayuki Takeda, and Ayumi Shinohara.
Ternary Directed Acyclic Word Graphs.
Theoretical Computer Science, 323(1-2):97-111, 2004.

[17] Hideo Bannai, Shunsuke Inenaga, Ayumi Shinohara, Masayuki Takeda, and Satoru Miyano.
Efficiently finding regulatory elements using correlation with gene expression. *Journal of Bioinformatics and Computational Biology*, 2(2):273-288, 2004.

[18] Shunsuke Inenaga, Ayumi Shinohara, Masayuki Takeda.
A fully compressed pattern matching algorithm for simple collage systems.
Int. J. Found. Comput. Sci. **16**(6):1155-1166, 2005.

[19] Hideo Bannai, Heikki Hyyro, Ayumi Shinohara, Masayuki Takeda, Kenta Nakai, and Satoru Miyano.
An $O(N^2)$ Algorithm for Discovering Optimal Boolean Pattern Pairs.
IEEE/ACM Transactions on Computational Biology and Bioinformatics 1(4):159-170, (2004).

[20] Shunsuke Inenaga, Ayumi Shinohara, Masayuki Takeda.
An Efficient Pattern Matching Algorithm on a Subclass of Context Free Grammars,
In *Proc. 8th International Conference on Developments in Language Theory* (DLT2004), pp. 225-236, December 2004.

[21] Heikki Hyyrö, Jun Takaba, Ayumi Shinohara, and Masayuki Takeda.
On Bit-Parallel Processing of Multibyte Text,
In *Proc. The First Asia Information Retrieval Symposium* (AIRS 2004), pp.289-300, October 2004.

[22] Shunsuke Inenaga, Hideo Bannai, Heikki Hyyrö, Ayumi Shinohara, Masayuki Takeda, Kenta Nakai, and Satoru Miyano,
Finding Optimal Pairs of Cooperative and Competing Patterns with Bounded Distance.
In *Proc. The 7th International Conference on Discovery Science* (DS 2004), pp. 32-46, October 2004.

[23] Hideo Bannai, Heikki Hyyrö, Ayumi Shinohara, Masayuki Takeda, Kenta Nakai, and Satoru Miyano
Finding Optimal Pairs of Patterns. *Proc. The 4th Workshop on Algorithms in Bioinformatics* (WABI 2004), pp. 450-452, September 2004.

[24] Shunsuke Inenaga, Ayumi Shinohara, and Masayuki Takeda.
A Fully Compressed Pattern Matching Algorithm for Simple Collage Systems.
In *Proc. The Prague Stringology Conference'04* (PSC'04), Czech Technical University Press, September 2004.

[25] Shunsuke Inenaga, Ayumi Shinohara, Masayuki Takeda, and Setsuo Arikawa.
Compact Directed Acyclic Word Graphs for a Sliding Window.
Journal of Discrete Algorithms , **2**(1):33-51, March 2004

[26] Toshiyuki Kochi, Akira Ishino, Masayuki Takeda, and Fumihiro Matsuo.

An unsupervised approach to syntactic ambiguity resolution using statistical information.
In *Proc. International Symposium on Information Science and Electrical Engineering 2003*, pp. 613-616, November 2003.

[27] Hisashi Tsuji, Akira Ishino, Masayuki Takeda, and Fumihiro Matsuo.
On space economical implementation of suffix trees.
In *Proc. International Symposium on Information Science and Electrical Engineering 2003*, pp. 203-206, November 2003.

[28] ○Masayuki Takeda, Shunsuke Inenaga, Hideo Bannai, Ayumi Shinohara, and Setsuo Arikawa.
Discovering Most Classificatory Patterns for Very Expressive Pattern Classes,
In Proc. 6th International Conference on Discovery Science (DS 2003), Lecture Notes in Computer Science 2843, pp. 486-493, October 2003.

[29] Tomohiko Sugimachi, Akira Ishino, Masayuki Takeda, and Fumihiro Matsuo.
A Method of Extracting Related Words Using Standardized Mutual Information.
In Proc. 6th International Conference on Discovery Science (DS 2003), Lecture Notes in Computer Science 2843, pp. 478-485, October 2003.

[30] Shunsuke Inenaga, Takashi Funamoto, Masayuki Takeda, and Ayumi Shinohara.
Linear-time off-line text compression by longest-first substitution.
In Proc. 10th International Symposium on String Processing and Information Retrieval (SPIRE 2003), Lecture Notes in Computer Science 2857, pp. 137-152, October 2003.

(2)口頭発表

①学会

国内　9　件,　　海外　0　件

[1] 御手洗 秀一，石野 明，竹田 正幸.
XML文書フィルタリングのための軽量な高速化技法.
電子情報通信学会第17回データ工学ワークショップ（DEWS2007), 2007.3.

[2] 池末 修也，石野 明，御手洗 秀一，竹田 正幸.
パスプルーニングと決定性有限オートマトンを用いた大規模かつ高速なXQuery処理システムの実装.
電子情報通信学会第17回データ工学ワークショップ（DEWS2006), 2006.3.

[3] 稲永 俊介，竹田 正幸.
Word Suffix Trees Revisited.
第61回人工知能基本問題研究会(SIG-FPAI), 2005.11.

[4] 坂内 英夫，畑埜 晃平，稲永 俊介，竹田 正幸.
Practical Algorithms for Pattern Based Linear Regression
第61回人工知能基本問題研究会(SIG-FPAI), 2005.11.

[5] 石野　明，竹田　正幸.
パスプルーニングと決定性有限オートマトンを用いたストリーム指向のXQuery処理.
データベースとWeb情報システムに関するシンポジウム(DBWeb2005), 2005.11

[6] 杉本 典子，金丸 玲子，関 隆宏，石野 明，竹田 正幸，廣川 佐千男.
XDES - 多様な構造と流動的変化に対応できるデータエントリーシステムの構築.
第 4 回情報科学技術フォーラム, 2005.

[7] 石野明，竹田正幸
ストリーム志向のXQuery処理系について，情報処理学会，デジタルドキュメント第49回研究会，2005.3.

[8] 田中 省作，杉本 典子，関 隆宏，石野 明，金丸 玲子，竹田 正幸，廣川 佐千男
大学経営における大学評価システムの活用，情報処理学会第67回全国大会, 2005.3.
[9] 杉本 典子，関 隆宏，石野 明，金丸 玲子，竹田 正幸，廣川 佐千男
XMLデータベースによる大学評価システムの構築，情報処理学会第67回全国大会，
2005.3.

②その他
国内　0　件，　海外　0　件

(3)特許出願（本研究に係わり、JST から出願したものとで研究機関から出願したもの）

出願元	国内（件数）	海外（件数）
JST		
研究機関		
計	なし	なし

(4)その他特記事項

とくになし

９．結び

本研究では知識発見基盤技術の開発について，理論と実用の両面から研究を行なっており，いずれについても満足のいく結果を得ることができた．本研究では，「非定型データ」を基本的に単なる文字の連鎖として扱っており，データの内容は問わない．したがって，Web ページなどの自然言語文はもとより，遺伝子情報，音楽情報などにも適用できるなど，応用範囲は極めて広い．

本研究で開発した技術により，計算資源(CPU パワーやメモリ・ディスク容量)の乏しい組込み機器であっても情報獲得・知識発見処理を行なうことが可能となる．来るべきユビキタス社会においては，この技術は，ユビキタス情報機器の組込みアプリケーションにとって不可欠な技術へとつながるものと期待できる．

本研究で得た研究成果をさらに発展させることにより，新世代型データベース基盤技術を構築する計画である．