研究課題別　事後評価結果

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基盤研究

２．研究代表者： 竹田　正幸　（九州大学　大学院システム情報科学研究院　教授）

３．研究内容及び成果

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。例えば、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、全体で一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。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”（公開は容易、質問は困難）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。

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望まれている。

このような背景のもとに、本研究は以下の3つの特色を有する。

①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とするが、文字の連鎖にはそれに相応しい処理方式をとる必要がある。本研究では、表面上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した。

②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出すことを言う。一方、情報科学の古典的研究分野の一つであるデータ圧縮も、データが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分かる通り、データ圧縮と機械発見の二つには密接な関係がある。そこで、機械発見の視点からデータ圧縮技術を捉え直すことで、こ

れまで見過ごされていた技術が脚光を浴びる可能性がある。したがって、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した。

③知識発見システムの成功の鍵は、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に目が行きがちであるが、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。さきがけ研究で、専門家の反応は設定した仮説空間の質に左右されることが示されている。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関与し得ない。一方、計算時間の面からは、仮説は単純であるほど良い。この相反する二つの要件の間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を解析した。

以下に示す3つの項目ごとに研究成果を挙げる。

1) 非定型データの高速パターン照合技術

研究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(Extensible Markup Language)ストリーム処理器XAXEN(eXtreamly-Accelerated XML filtering ENgine)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あり、ストリーム処理手法として有名なXMLTK(XML Toolkit for Scalable XML Stream Processing)やYFilter(XML filtering system)と比較して実行速度で約4～6倍、メモリ使用量で約6分の1以下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、XAXENが圧倒的に優れていることが明らかとなった。

2) 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一つの技術として、研究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、例えば英文テキストの場合、従来手法では60%程度しか実行時間を短縮出来なかったものを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の文法サイズ最小化問題がNP(Non-deterministic Polynomial time)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(schema)に着目した。この圧縮は、初歩的な方法では$O(n^4)$時間を要し、Minimal Augmented Suffix Treeと呼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。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実行時間を劇的に削減し、$O(n)$時間で計算するアルゴリズムを開発した。

3)非定型データからのパターン発見技術

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するとともに高速化するためのデータ構造開発を行った。応用として、音楽データ・言語データ・薬学データ・税関の申告データ等様々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、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。

4. 事後評価結果

4－1. 外部発表(論文、口頭発表等)、特許、研究を通じての新たな知見の取得等の研究成果の状況

期間中の外部発表、特許等の実績を示す。

発表論文: 30 件

口頭発表: 9 件

非定型データの高速パターン照合ソフトウェアを開発し、商用ソフトよりもはるかに性能の良いソフトウェアを実現した。先進的なアルゴリズム開発とソフトウェアとしての実現を併せて実施した点が評価出来る。パストライというデータ構造を利用したXMLテキストサーチアルゴリズムの提案と、既存のシステムに比べて実行速度やメモリ効率の点で優れたXAXENシステムの実現、さらに医薬品や文学等への応用を行い、十分な成果を上げたと思われる。ただし、知識発見とは単にパターンを発見することではなく、その意味を推論することも含まれているはずであるが、それについての成果は明確ではない。英語で28件の論文が発表されており、量的には問題ない。また、この分野の主要会議での発表や主要雑誌での掲載等、質的にも十分である。CiteSeerによる引用数も10を超える論文があり、世界的に見ても影響力のある

成果である。

４－２．成果の戦略目標・科学技術への貢献

種々のテキスト処理システムで適用可能な技術を実現しているとともに、非テキスト情報への応用の可能性もあり、強いインパクトがある。XMLを対象としたパターン照合については我が国随一の研究グループに成長しており、今後に期待出来る。開発されたアルゴリズムやソフトウェアを用いて、医薬品商標名の類似性と投薬ミスの解析、通関データからの不正輸出の検出、Webスパムの検出等の実用的な問題に対応する技術を開発していることは高く評価される。今回の成果を一層洗練させていくことは極めて重要であるが、これをベースにした新たな展開の方向性が明確ではないため、応用への適用可能性等を十分に検討して今後の目標を設定することが必要である。特に、情報洪水時代に真剣に対処しようとするならば、半構造テキストデータのみを対象にするのは明らかに不十分であり、マルチメディアデータ等の非テキストパターンデータを含め、高速に検索・照合する仕組みが必要だと思われる。今後は、若手育成への配慮、研究拠点形成への努力、種々の応用事例の開発等が並行して行われることを期待する。