研究課題別研究評価

1. 研究課題名 :最適パターン発見にもとづく高速テキストデータマイニング

2. 研究者名 :有村博紀

3. 研究の狙い

本研究の目的は, ネットワーク上に蓄積された膨大なテキストと半構造データから, 有用な情報を獲得するための高速なテキストデータマイニングシステムを開発することである。ウェブページやXMLデータ等のネットワーク上の大規模テキストデータの利用が急速に進みつつある現在, テキストデータからの効率良いデータマイニング手法の確立が緊急の課題となっている。しかしその一方で, これらの大規模テキストデータは, (1) 明示的な構造をもたない, (2) 多様な電子化文書の, (3) 膨大な量の集積であるという特徴をもっており, 関係データベースを対象に開発されてきた従来型のデータマイニング技術をそのまま適用することができないという問題があった。そこで本研究では, 従来型技術の活用ではなく, 新しい観点からテキストマイニングの問題に正面から取り組み, テキストデータマイニングのための基本技術の研究開発を行なう。さらに, これら基本技術の開発を通じて, 大規模テキストデータを対象とした高速テキストマイニングシステムのプロトタイプを構築する。研究の特色として, データマイニングを, 人間による大量のデータ解析を支援する効率的な半自動的ツールとしてとらえ, 従来の情報検索システムを超えた新しい情報アクセスシステムの開発を目指す。また, 計算量理論と計算学習理論との最新の成果を積極的に取り入れて, 大量のデータに対してきわめて高速かつ頑健に働くアルゴリズムの開発を目標とする。

4. 研究結果

本研究では, 大規模テキストデータから, データを特徴付けるパターンを高速に発見するための, 一連の高速なテキストデータマイニング手法を開発し, ネットワーク上の大規模テキストデータを対象としたテキストデータマイニングシステムの実現方式を明らかにする。本研究構想のポイントは, (1) 最適パターン発見の枠組みに基づく新しいテキストマイニングの枠組みの提案と, この枠組みに基づく一連の高速な最適パターン発見アルゴリズムの開発, (2)テキストマイニングシステムの大規模実装のための基盤技術の開発, (3)ウェブからの情報抽出と半構造データマイニングへの拡張, (4)提案の枠組みと開発した技術の有効性を検証するための大規模テキストマイニング実験である。

さきがけ研究の3年間において, 理論・実装・応用の3つの観点からこれらのポイントに重点を置いて研究を行ない, 以下の研究結果を得た。

(1) 最適パターン発見に基づく高速な最適パターン発見アルゴリズムの開発。

- テキストデータマイニングを実現するための新しい枠組みとして最適パターン発見の枠組みを提案した。さらに, この最適パターン発見に基づいて, 文字列アルゴリズムと計算幾何学研究の最新の成果を援用して, 大量のテキストデータを特徴づけるパターンを高速に発見するテキストマイニングツール AWAP(Algorithm for Word- Association Patterns)を開発した。最適パターン発見は, 多様なデータに対する頑健性と有用性が認識されている反面, 高い計算

量をもつため，テキストデータに対する実用的なアルゴリズムの実現は難しいと考えられていた。今回の結果により，大規模テキストマイニングにおいても効率良い最適パターン発見が可能であることを示し，高速かつ頑健なテキストマイニングを実現するための基礎技術を確立することができた。

(2) テキストマイニングシステムの大規模実装のための基盤技術の開発。

1. 従来テキスト解析に利用されてきた接尾辞木(suffix tree)索引構造の代わりに，接尾辞配列(suffix array)と高さ配列(height array)という1 次元整数配列を組み合わせて，大規模化可能で記憶効率が良いマイニング向けテキスト索引技法を開発した。このために必要な技術として，接尾辞配列の一方向走査を用いた接尾辞木の巡回の模倣，索引構造の動的な再構成法，当時知られていなかった接尾辞配列からの高さ配列の線形時間構築法等，高速テキストマイニング実現のためのさまざまな基盤技術を新たに開発した。これらの技術開発により，AWAP アルゴリズムの 2 桁から3 桁以上の高速化を達成した。
2. 大規模データマイニングアルゴリズムに用いられるディスク走査技法に基づいて，AWAP と相補的な性能をもち，外部記憶におかれた大規模テキストを扱うことのできるような，高速な最適パターン発見アルゴリズム LevelwiseScan を開発した。この技術により，実用的な計算資源だけを用いて，数百メガバイトを超えるような大量のテキストデータを対象としたテキストマイニングが可能になった。

(3) ウェブからの情報抽出と半構造データマイニング。

1. ウェブページやXMLデータ等の構造をもったテキスト(半構造データ)からのパターン発見問題を考察し，大量の半構造データからの頻出する部分構造を発見するアルゴリズム FREQT を開発した。さらに，これを最適パターン発見の枠組みに拡張し，情報エントロピーや分類精度等の統計的尺度に関する最適部分構造を高速に発見するアルゴリズム OPTT を与えた。また，理論的解析と実データを用いた計算機実験によって，ウェブマイニングにおける提案手法の有効性を実証した。これらの結果により，今後の半構造データマイニングのための基本的技術を確立した。
2. ウェブデータからの情報抽出問題を考察し，与えられたウェブページとそこからの切り出し例から，ウェブページの構造情報を用いて高精度な情報抽出を行なうラッパー構築手法 TreeWrapper と一連のアルゴリズムを開発した。実際のウェブページを用いた計算機実験では，従来手法に比較して精度の高い抽出結果が得られた。

(4) テキストマイニングシステムのプロトタイプ構築と応用実験。

1. 開発したアルゴリズムとデータ構造を基に，テキストマイニングのプロトタイプシステムを構築し，大規模テキストデータを用いたテキストマイニング実験を行なった。英文新聞記事データからの探索的文書ブラウジング実験と，ウェブ検索エンジンおよびウェブロボットを用いたネットワーク上のウェブページからのキーワード獲得実験を行い，最適パターン発見の有効性を示した。これにより，最適化パターン発見をもちいることで，ウェブデータのような多様な質と内容をもつ大規模テキストデータに対して，自然言語処理や領域依存の統計的経験則を用いることなく，有用なパターン発見が可能であることを示した。

5. 自己評価

さきがけ研究での最大の目標であった，高速かつ頑健なテキストマイニングシステムについて

は, そのためのテキストマイニングへの最適パターン発見の導入および, 高速なパターン発見技術である AWAP と一連のアルゴリズム, 大規模テキスト索引技術, XMLデータ等の半構造データに対する拡張により, 大規模テキストマイニングのための基本的技術を開発することができた。さらに, プロトタイプシステムによる実験においては, 大規模テキストデータの探索的ブラウジングや, ウェブからの キーワード発見等の実データを用いた応用実験によって, ネットワーク上の大規模データに対する情報獲得において, 提案技術の応用可能性を示すことができたことも, 本研究の有効な成果であったと考える。これらの研究成果により, 最適パターン発見を用いたテキストマイニングが, 従来型の情報アクセス技術を補完する新しい情報アクセス手段として有効性をもつことを示せたと考える。

このように, 最適パターン発見に基づいたテキストマイニングと, そのための基盤技術の研究開発については, 当初予定していた以上に研究を進めることができ, 基本的技術を確立できたと考える。一方で, さきがけ研究開始時点での目標であった「並列テキストマイニングによる大規模テキストマイニング・システムの実現」に関しては, マイニング向け大規模テキスト索引の並列化を行った段階であり, 完全な並列テキストマイニングシステムの構築までにはいたらなかった。

今後, この最適パターン発見技術を, ネットワーク上にあふれる膨大な量の情報から, 自分が必要とする情報を得るための柔軟かつパワフルな道具とするためには, さきがけ研究の中心にすえてきたような高速なアルゴリズムの開発だけでなく, 人間の知識獲得活動支援にこれらの技術をどのように埋め込むべきかといったシステム設計と人的因子等の問題も考えていく必要があると考えている。今後は, 他分野・対象領域の専門家との協同研究も進めながら, 研究を進めていきたい。

半構造データマイニングについては, 現在急速に進展している分野であり, 現在, 半構造データの基本的問題の一部分を解決できたという段階である。今後は, さきがけ研究で遂行したテキストマイニングの研究を, 半構造データマイニング研究にどのように展開させていくべきかさらに探求をする必要がある。さきがけ終了後もさらに研究を遂行中であるが, 現在, データマイニング分野では半構造データマイニングに関して, さまざまな技術提案や応用可能性が出されており, この方向にも, さらに継続して研究を進めたい。

大規模テキストデータからの知識獲得は, 急速に情報社会化が進み, ネットワーク上に大量の情報があふれる中で, 個人や組織が必要な情報を主体的に獲得し, 迅速な決定を可能とするための鍵となる技術である。新しい型の大規模データであるXMLデータ等の半構造データ技術の急速な発展と普及は, このような情報獲得技術の重要性をますます強めるものである。本研究は, 従来型技術の単なる組み合わせやアドホックな方法で一時的な解決を求めるのではなく, 新しいアプローチで, 大規模データに対する新しいアクセス技術を実現しようとするものである。

さきがけ研究の 3 年間では, 本研究構想の実現に向けた基本的な研究成果を得ることができた。研究開発した技術が広く使われる有効な技術となるためには, 今後も多くの問題を解決する必要があるが, 理論にもとづいた応用と, 応用に深く根ざした理論という正統的な計算機科学の方法論にもとづいて, 最終的な目標に近づいていきたい。さきがけ研究の 3年間では, 研究の基本構想の提案から, 基礎技術の開発, 応用への展望に関して, 一貫した研究活動を行なうことができ, 本研究構想の実現において大きく前進することができた。3 年間のさきがけ研究による支援に多謝するとともに, 今後も研究を発展させていきたいと考えている。

6. 研究総括の見解

緊急の課題となっているネットワーク上の大規模データからの知識獲得の技術に関して、膨大、多様で高速かつ頑健なテキストマイニングのための新しい基盤技術の確立を目指し、たゆまぬ努力により顕著な成果をあげた。具体的には、高速な最適パターン発見アルゴリズムの開発、テキストマイニングの大規模実装のための基盤技術の開発、また、大量の半構造データマイニングからの頻出する部分構造を発見するアルゴリズムの開発やプロトタイプの構築などがあるが、今後の研究発展が強く期待できる。

7. 主な論文等

論文〕

- T. Asai, H. Arimura, K. Abe, S. Kawasoe, and S. Arikawa, Online Algorithms for Mining Semi-structured Data Stream, Proc. IEEE International Conference on Data Mining (ICDM'02), IEEE Computer Sciety Press, December 2002.
- K. Abe, S. Kawasoe, T. Asai, H. Arimura, S. Arikawa, Optimized Substructure Discovery for Semi-structured DataProc. 6th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD-2002), LNAI 2431, Springer-Verlag, 1-14, 2002.
- T. Asai, K. Abe, S. Kawasoe, H. Arimura, H. Sakamoto, S. Arikawa, Efficient Substructure Discovery from Large Semi-structured Data, Proc. Second SIAM International Conference on Data Mining (SDM'02), 158-174, SIAM, 2002.
- H. Sakamoto, K. Hirata, and H. Arimura, Learning Elementary Formal Systems with Queries, Theoretical Computer Science, 2002. (accepted)
- 村上義継、坂本比呂志、有村博紀、有川節夫（九大）, HTML からのテキストの自動切りだしアルゴリズムと実装, 情報処理学会論文誌 :数理モデル化と応用、Vol. 42, No. SIG 14 (TOM 5), 39-49, Dec 2001.
- 安積裕樹、川副真治、安部潤一郎、有村博紀、有川節夫（九大）、分散記憶型並列計算機における大規模接尾辞配列の構築法, 情報処理学会論文誌 :数理モデル化と応用、Vol. 42, No. SIG 14 (TOM 5), 14-24, Dec 2001.
- H. Arimura, H. Sakamoto, S. Arikawa, Efficient Learning of Semi-structured Data from Queries, Proc. the 12th International Conference on Algorithmic Learning Theory (ALT'01), LNAI 2225, 315-331, Springer-Verlag, 2001.
- K. Taniguchi, H. Sakamoto, H. Arimura, S. Shimozono and S. Arikawa, Mining Semi-Structured Data by Path Expressions, Proc. the 4th International Conference on Discovery Science, LNAI 2226, 378-388, Springer-Verlag, 2001.
- A. Yamamoto, K. Ito, A. Ishino, H. Arimura, Proc. the 11th International Conference on Inductive Logic Programming (ILP'01), LNAI 2157, Springer-Verlag, 2001
- T. Kasai, G. Lee, H. Arimura, S. Arikawa, K. Park, Linear-time Longest-Common-Prefix Computation in Suffix Arrays and Its Applications, Proc. the 12th Annual Symposium on Combinatorial Pattern Matching (CPM'01), LNCS 2089, 181-192, Springer-Verlag, 2001.
- H. Arimura, H. Asaka, H. Sakamoto, S. Arikawa, Efficient Discovery of Proximity Patterns

with Suffix Arrays (Extended Abstract), Proc. the 12th Annual Symposium on Combinatorial Pattern Matching (CPM'01), Short talk, LNCS 2089, 152-156, Springer-Verlag, 2001.
- H. Sakamoto, H. Arimura, and S. Arikawa, Extracting Partial Structures from HTML Documents, Proc. the 14th Florida Artificial Intelligence Research Symposium (FLAIRS'2001), Florida, AAAI, 264-268, May, 2001.
- H. Arimura and S. Jain (eds.), Proc. the 11th International Workshop on Algorithmic Learning Theory (ALT'00), LNAI 1968, Springer-Verlag, Sydney, Dec. 2000。
- H. Arimura, J. Abe, R. Fujino, H. Sakamoto, S. Shimozono, S. Arikawa, Text Data Mining: Discovery of Important Keywords in the Cyberspace , Proc. Kyoto International Conference on Digital Libraries 2000, Kyoto University, British Library and National Science Foundation (U.S.A.), 121-126, 2000.
- H. Sakamoto, H. Arimura, S. Arikawa, Identification of Tree Translation Rules from Examples, Proc. the 5th International Colloquium on Grammatical Inference (ICGI 2000), LNAI 1891, Springer-Verlag, 241-255, Sep. 2000.
- 安部 潤一郎, 藤野 亮一, 下薗 真一, 有村 博紀, 有川 節夫、テキストデータからの高速データマイニング人工知能学会誌, Vol.15, No.4, 2000 年 7 月
- H. Arimura, H. Sakamoto, and S. Arikawa, Learning Term Rewriting Systems from Entailment, 10th International Conference on Inductive Logic Programming (ILP2000) Work-in-Progress paper session, July 2000.
- H. Arimura, Text Data Mining with Optimized Pattern Discovery, Proc. the 17th Machine Intelligence - Life Long Learning and Discovery in Procedural and Declarative Knowledge, K. Furukawa, S. Muggleton, D. Michie, and L. DeRaedt (eds.), 2000.
- R. Fujino, H. Arimura, S. Arikawa, Discovering Unordered and Ordered Phrase Association Patterns for Text Mining, Proc. 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD2000), LNAI 1805, 281-293, Springer-Verlag, Nara, Apr. 2000.
- S. Shimozono, H. Arimura, and S. Arikawa, Efficient Discovery of Optimal Word-Association Patterns in Large Text Databases, New Generation Computing, 18, 49 - 60, 2000.
- A. Yamamoto and H. Arimura, Inductive Logic Programming : From Logic of Discovery to Machine Learning, Special Issue on Surveys on Discovery Science, (Eds.) S. Miyano, IEICE Transaction on Infomation and System, E83-D (1), 10-18, 2000.

解説記事〕
- 有村 博紀, 坂本比呂志、データマイニングにおける最適パターン発見、応用数理、応用数理学会, 第 12 巻第 4 号, 2002.
- 池田 大輔・坂本 比呂志・有村 博紀, ウェブデータマイニング, システム/制御/情報 「データマイニング特集号」, システム制御情報学会, 第 46 巻第 4 号, Apr. 2002.
- 坂本比呂志、有村博紀、Web マイニング, 特集 「テキストマイニング」、人工知能学会誌, Vol. 16, No. 2, 2001 年 3月.
- 那須川哲哉、河野浩之、有村博紀、テキストマイニング基盤技術, 特集 「テキストマイニング」、

人工知能学会誌, Vol. 16, No. 2, 2001 年 3月.

受賞〕

- 電子情報通信学会 DE 研究、DEWS2002 優秀論文賞、2002 年 5月受賞。
- 人工知能学会 2000 年度論文賞, 2001 年 5 月受賞。
- PAKDD2000 Paper with Merit Award, 2000 年 4 月受賞。
- 人工知能学会 1999 年度全国大会優秀論文賞、1999 年 12 月受賞。

招待・依頼講演〕

- H. Arimura, Efficient Text Mining with Optimized Pattern Discovery (invited talk), Proc. the 13th Annual Symposium on Combinatorial Pattern Matching (CPM'02), LNCS 2373、17-19, Springer-Verlag, Fukuoka, July 2001. 招待講演）
- 坂本比呂志、村上義継、安部潤一郎、有村博紀、有川節夫、ウェブからの情報抽出と最適パターン発見, 特別セッション「データ・テキストマイニングにおける統計的モデリングの実際」、第 4 回情報論的学習理論ワークショップ（IBIS2001）、117-122, 2001. 招待講演）
- 有村博紀、最適パターン発見にもとづくデータマイニング, 統計数理とデータマイニング、統計数理研究所共同研究リポート、142、13-24, 統数研、2001. 依頼講演）
- 有村博紀、データマイニング－ ウェブデータからの 知識発見を目指して－, 電子情報通信学会 IT 研究会、2000. 招待講演）
- チュートリアル企画、 発見科学とデータマイニングの最前線 :金融・経済・ゲノム からウェブまで」、2001 年電子情報通信学会ソサイエティ大会、2001 年 9 月. 依頼講演）
- 有村博紀、テキストマイニング :ウェブデータからの 知識発見を目指して、第２5回情報化学討論会概要集, J13, 日本化学会情報化学部会, 2002. 招待講演）