研究課題別評価

1. 研究課題名 :文字列データ圧縮に基づく高速知識発見システムの構築

2. 研究者氏名 :竹田　正幸

3. 研究の狙い：

大量のデータから、そこに内在する規則や傾向などを、計算機によって半自動的に発見する「機械発見」技術の確立が強く望まれている。本研究では、データ圧縮という古典的研究分野に「機械発見処理の高速化」という新しい価値基準を導入し、この視点から、データ圧縮で用いられる各要素技術の再評価を行い、機械発見システム構築のための基礎技術を確立することを目的とする。

本研究では、対象を、陽には構造をもたない文字列データに絞り、文字列データを対象とした機械発見の問題を扱う。データ圧縮と機械発見を統一的に扱うために、文字列記述の形式的体系を導入し、このもとで、機械発見に必要な文字列処理問題の計算量を解析してその階層を究明する。また、実用的に有用なクラスに関して、文字列照合や文字列データ圧縮の技術を駆使した高速なアルゴリズムを開発することを目指す。

4. 研究結果：

文字列データからの知識発見のための要素技術の確立を目指して、理論と実用の両面から研究に取り組んできた。その成果の概要は以下の通りである。

【理論的側面】

[1] 本研究の第 1 の特色は、データ圧縮を処理速度向上の手段として用いるという、一見常識に反する手法を提案し、高速化に成功した点である。圧縮による高速化の研究は、フランスのマルネラ・ラ・バレー大学の M. クロシュモア教授、イスラエルの G. ランダウ教授、米国のブランディス大学 J. ストーラー教授など、当該分野第一線の研究者から高く評価され、このアイデアに追随した研究も多く見受けられるようになった。圧縮による高速化に関して、以下の成果を得た。

1. 複数パターン照合の高速化。
2. 近似文字列照合の高速化。

[2] 本研究の第 2 の特色として、従来ディスク容量の節約やデータ転送量の低減のみを目的としていたデータ圧縮を、機械発見の観点から捉え直した点が挙げられる。ここでは、「パターンによる圧縮」という概念を導入することによって、「パターン発見」と「テキスト圧縮」という一見まったく別の問題が、実は密接に関連していることを顕在化させた。具体的には、以下の成果が得られた。

1. 部分文字列パターン、部分列パターン、VLDC パターン、Angluin パターン、断片パターンなど、様々な形式のパターンのクラスに対し、
   (a) パターン照合問題、
   (b) 最適弁別パターン発見問題、

(c) 最適パターン発見問題、
(d) 類似度計算問題
の各々についての計算量階層の究明。

2. 1 で多項式時間計算可能と判明したパターンクラスについての高速アルゴリズムの開発。
3. 1 で計算困難性が示されたものについて、実用時間内に動作させるための有効な枝刈り方式等の開発。

[3] 以下で述べる「実用的研究」の[1]、[2]で得た知見を生かし、ウインドウ幅の制限や近似的照合を加味したパターン照合を盛り込んだ、より表現力の豊かなパターンを対象に、高速パターン発見処理を理論的に追求した。具体的には、以下の 2 つについて研究を進め、効率的手法の開発に成功した。

1. 表現力のアップしたパターンのクラスについての計算量階層究明。
2. 有効な枝刈り手法の開発。
3. さらなる高速化のための、高速パターン照合技術の開発。

【実用的側面】

[1] 知識発見システムにおいては、システムは、あらかじめ設定した仮説空間の中から最もスコアの高い仮説を選び出し出力する。だが、この「仮説」がそのまま有益な発見につながるわけではない。専門家による仮説の評価・意味づけ作業を通して、初めて有益な知見が得られる。この評価は、必ずしも体系的なものではなく、自然科学分野においてすら、「学問的勘」や長年の経験に基づく「コツ」と無縁ではない。したがって、成功のカギは、このような専門家の「主観的判断」をどれだけうまくシステムに取り込み得るかにかかっている。

そこで、研究の第 1段階においては、上述の理論的研究の傍ら、最も主観的判断を伴う学問領域のひとつとして文学を選び、文学研究者の協力を得て「古典和歌からの知識発見」という研究テーマに挑んだ。その結果、国文学上の研究成果として学会発表され学術雑誌論文にもまとめられるほどの、レベルの高い発見がなされ、3 度にわたって新聞報道されるなど、社会的にも大きなインパクトを与えた。

[2] この経験を通じて、専門家の主観的知見をも取り込むためのノウハウを蓄積することができた。すなわち、専門家の問題意識や対象領域の性質をできる限り反映した、柔軟でより表現力のある仮説空間を設定することの重要性が判明した。一方、計算量の観点からは、できるだけ単純な仮説空間を設定しておかなければ、実用的時間内に計算を終了させることが困難である。そこで、研究の第 2段階においては、より豊かな仮説空間に対する知識発見の問題を可能な限り高速に処理するための研究に、真正面から取り組んだ。かなり難易度の高い課題であった、膨大な探索空間に対する有効な枝刈り方式と、複雑なパターンに対する高速照合技術の開発に成功した。

また、いわゆるデータ発掘の出力は、そのままで有用であることはなく、人間による評価・解釈作業が重要となるが、この作業を支援する方策として、パターンのコンパクトな表現法に基づく可視化の研究も併せて行い、ソフトウェアを作成し実用に供した。

5. 自己評価：

本研究では、テキストデータからの知識発見の核となる要素技術の開発を目指した。その最大の目的は、知識発見処理の高速化にあった。だが、誰も使わないような処理を高速化しても無意味である。このような研究を有意義なものとするためには、現実社会に存する問題をうまく定式化することにより、真に意味のある知識発見問題について高速化を目指すべきである。そこで、単にアルゴリズムを開発して処理技術の高速化を図るだけでなく、一方で現実の問題へ適用し、本研究で開発した知識発見処理の有効性を検証し続ける必要がある。すなわち、理論と応用の両面から、知識発見の問題へ取り組む必要がある。

本研究では、このような観点から、3年間知識発見の研究に取り組み、満足のいく成果をあげることができた。ここで扱っているテキストデータとは、文字の連鎖であればよく、内容は問わない。したがって、本研究で得られた要素技術は、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、応用範囲はきわめて広い。

最適弁別パターン発見問題に関しては、本研究で開発した技術を使って、DNA の塩基配列やアミノ酸配列を対象とした実験を、分子生物学者の協力を得ながら引き続き行っている。仮説空間を変更した後の分子生物学者側の反応は、これまでと明らかに違っており、生物学上の有益な発見が得られるものと期待できる。

また、類似度計算問題に関しては、音楽情報の他、医薬品取り違え防止の観点から、薬名間類似性指標の設計を目指し、薬学分野の研究者との共同研究を遂行中である。

さらに、本研究の成果として得られた技術は、時系列データ解析にも有効と思われる。現在、いくつかの具体的な問題に本手法を適用し、その解決に取り組んでいる。

6. 研究総括の見解：

急速に情報社会化が進みネットワーク上に情報が氾濫する中で、個人や組織が必要な情報を主体的に獲得し、迅速な意思決定を行うために不可欠である「大量データからの知識発見技術の開発」という緊急の課題に取り組み、大量・不定形・不均質なデータから知識獲得を行うための新しい基盤技術の確立を目指し、顕著な成果をあげた。具体的には、特徴パターン発見，最適弁別パターン発見，類似文字列発見という3つの問題のそれぞれに対し、多様な仮説空間に対する高速な発見アルゴリズムを開発・実装し、発見支援システムに結実させた。また、理論と実装にとどまらず、応用面においても、言語学や文学，分子生物学といった他分野の研究者と手を組み、現場の具体的な問題に対して開発手法の有効性を示した点は、高く評価できる。今後の研究発展が強く期待される。

7. 主な論文等

論文（査読付き論文のみ）

1. M. Takeda, S. Inenaga, H. Bannnai, A. Shinohara and S. Arikawa: Discovering Most Classificatory Patterns for Very Expressive Pattern Classes. In *Proc. 6th International Symposium on Discovery Science (DS2003)*, 486-493, Springer-Verlag, October 2003.
2. S. Inenaga, T. Funamoto, M. Takeda and A. Shinohara: Linear-Time Off-Line Text

Compression by Longest-First Substitution. In *Proc. 10th International Symposium on String Processing and Information Retrieval (SPIRE2003)*, 137-152, Springer-Verlag, October 2003.

3. K. Baba, S. Tsuruta, A. Shinohara, and M. Takeda: On the Length of the Minimum Solution of Word Equations in One Variable. In *Proc. 28th International Symposium on Mathematical Foundations of Computer Science (MFCS2003)*, 189--197, Springer-Verlag, August 2003.
4. K. Baba, A. Shinohara, M. Takeda, S. Inenaga, and S. Arikawa: A Note on Randomized Algorithm for String Matching with Mismatches. *Nordic Journal of Computing* 10:2-10 (2003).
5. H. Bannai, S. Inenaga, A. Shinohara, and M. Takeda: Inferring Strings from Graphs and Arrays. In *Proc. 28th International Symposium on Mathematical Foundations of Computer Science (MFCS2003)*, 208-217, Springer-Verlag, August 2003.
6. T. Kida, T. Matsumoto, Y. Shibata, M. Takeda, A. Shinohara, and S. Arikawa: Collage system: A unifying framework for compressed pattern matching. *Theoretical Computer Science* 298(1):253-272 (2003).
7. S. Miyamoto, S. Inenaga, M. Takeda, and A. Shinohara: Ternary Directed Acyclic Word Graphs. In *Proc. Eighth International Conference on Implementation and Application of Automata (CIAA2003)*, Springer-Verlag, July 2003.
8. M. Hirao, H. Hoshino, A. Shinohara, M. Takeda, and S. Arikawa: A practical algorithm to find the best subsequences patterns. *Theoretical Computer Science* 292(2):465-479 (January 2003).
9. M. Takeda, T. Fukuda, I. Nanri, M. Yamasaki, and K. Tamari: Discovering instances of poetic allusion from anthologies of classical Japanese poems. *Theoretical Computer Science* 292(2):497-524 (January 2003).
10. M. Takeda, T. Matsumoto, T. Fukuda, and I. Nanri: Discovering characteristic expressions in literary works. *Theoretical Computer Science* 292(2):525-546 (January 2003).
11. H. Bannai, S. Inenaga, A. Shinohara, M. Takeda, and S. Miyano: A String Pattern Regression Algorithm and Its Application to Pattern Discovery in Long Introns. In *Genome Informatics 13, (GIW2002)*, pp. 3-11, Universal Academy Press, Inc., December 2002.
12. M. Takeda, T. Fukuda, and I. Nanri: Mining from Literary Texts: Pattern Discovery and Similarity Computation. In *Progress in Discovery Science* (Final Report of the Japanese Discovery Science), Lecture Notes in Artificial Intelligence (LNAI2281), Springer-Verlag, 2002.
13. A. Shinohara, M. Takeda, S. Arikawa, M. Hirao, H. Hoshino, and S. Inenaga: Finding Best Patterns Practically. In *Progress in Discovery Science* (Final Report of the Japanese Discovery Science), Lecture Notes in Artificial Intelligence (LNAI2281), 307-317, Springer-Verlag, 2002.
14. S. Inenaga, H. Bannai, A. Shinohara, M. Takeda, and S. Arikawa, Discovering Best Variable-Length- Don't-Care Patterns. In *Proc. 5th International Conference on Discovery Science (DS2002), Lecture Notes in Artificial Intelligence (LNAI)*, Springer-Verlag, November

2002.

15. M. Takeda, S. Miyamoto, T. Kida, A. Shinohara, S. Fukamachi, T. Shinohara, and S. Arikawa: Processing text files as is: Pattern Matching over compressed texts, multi-byte character texts, and semi-structured texts. In *Proc. 9th International Symposium on String Processing and Information Retrieval (SPIRE2002)*, 170--186, Springer-Verlag, September 2002.
16. S. Inenaga, A. Shinohara, M. Takeda, and S. Arikawa: Compact Directed Acyclic Word Graphs for a Sliding Window. In *Proc. 9th International Symposium on String Processing and Information Retrieval (SPIRE2002)*, 310--324, Springer-Verlag, September 2002.
17. S. Inenaga, A. Shinohara, M. Takeda, H. Bannai, and S. Arikawa: Space-Economical Construction of Index Structures for All Suffixes of a String. In *Proc. 27th Inter. Symp. on Mathematical Foundation of Computer Science (MFCS2002)*, 341--352, Springer-Verlag, August 2002.
18. S. Inenaga, M. Takeda, A. Shinohara, H. Hoshino, and S. Arikawa: The Minimum DAWG for All Suffixes of a String and Its Applications. In *Proc. 13th Annual Symposium on Combinatorial Pattern Matching (CPM'02)*, 153--167, Springer-Verlag, July 2002.
19. H. Bannai, K. Iida, A. Shinohara, M. Takeda, and S. Miyano: More speed and more pattern variations for knowledge discovery system BONSAI. In *Genome Informatics 12 (GIW2001)*, 454--455, Universal Academy Press, Inc., December 2001.
20. H. Hori, S. Shimozono, M. Takeda, and A. Shinohara: Fragmentary pattern matching: Complexity, algorithms and applications for analyzing classic literary works. In *Proc. 12th Annual International Symposium on Algorithms and Computation (ISAAC'01)*, 719? 730, Springer-Verlag, December 2001.
21. M. Hirao, S. Inenaga, A. Shinohara, M. Takeda, and S. Arikawa: A practical algorithm to find the best episode patterns. In *Proc. 4th International Conference on Discovery Science (DS 2001)*, 432--437, Springer-Verlag, November 2001.
22. K. Yamamoto, M. Takeda, A. Shinohara, T. Fukuda, and I. Nanri: Discovering repetitive expressions and affinities from anthologies of classical Japanese poems. In *Proc. 4th International Conference on Discovery Science (DS 2001)*, 413--425, Springer-Verlag, November 2001.
23. T. Kadota, M. Hirao, A. Ishino, M. Takeda, A. Shinohara, and F. Matsuo: Musical sequence comparison for melodic and rhythmic similarities. In *Proc. 8th International Symposium on String Processing and Information Retrieval (SPIRE 2001)*, 111--122, IEEE Computer Society, November 2001.
24. S. Inenaga, H. Hoshino, A. Shinohara, M. Takeda, and S. Arikawa: On-line construction of symmetric compact directed acyclic word graphs. In *Proc. 8th International Symposium on String Processing and Information Retrieval (SPIRE 2001)*, IEEE Computer Society, 96? 110, November 2001.
25. S. Inenaga, H. Hoshino, A. Shinohara, M. Takeda, and S. Arikawa: Construction of the CDAWG for a trie. In *Proc. Prague Stringology Club Workshop (PSC2001)*, 37--48, Prague, Czech

Republic, September 2001.
26. T. Kida, T. Matsumoto, M. Takeda, A. Shinohara, and S. Arikawa: Multiple pattern matching algorithms on collage system. In *Proc. 12th Annual Symposium on Combinatorial Pattern Matching (CPM'01)*, 193--206, Springer-Verlag, July 2001.
27. M. Takeda: String resemblance system - A unifying framework for string similarity with applications to literature and music. In *Proc. 12th Annual Symposium on Combinatorial Pattern Matching (CPM'01)*, 147--151, Springer-Verlag, July 2001.
28. S. Inenaga, M. Hoshino, A. Shinohara, M. Takeda, S. Arikawa, G. Mauri, and G. Pavesi: On-line construction of compact directed acyclic word graphs. In *Proc. 12th Annual Symposium on Combinatorial Pattern Matching (CPM'01)*, 169--180, Springer-Verlag, July 2001.
29. M. Takeda, Y. Shibata, T. Matsumoto, T. Kida, A. Shinohara, S. Fukamachi, T. Shinohara, and S. Arikawa: Speeding up string pattern matching by text compression: The dawn of a new era. 情報処理学会論文誌 42(3): 370--384 (March 2001).
30. G. Navarro, T. Kida, M. Takeda, A. Shinohara, and S. Arikawa: Faster approximate string matching over compressed text. In *Proc. Data Compression Conference (DCC 2001)*, 459-468, IEEE Computer Society, March 2001.
31. S. Mitarai, M. Hirao, T. Matsumoto, A. Shinohara, M. Takeda, and S. Arikawa: Compressed pattern matching for SEQUITUR. In *Proc. Data Compression Conference (DCC 2001)*, 469-478, IEEE Computer Society, March 2001.
32. 竹田正幸, 福田智子, 南里一郎, 山崎真由美, 玉利公一 和歌データからの類似歌発見. 統計数理 48(2): 289--310 (December 2000).

解説記事（学術雑誌）
1. 竹田正幸: コンピュータは文学研究を変えるか？人工知能学会誌 17(3):326-330, 2002.
2. 竹田正幸, 篠原歩: 圧縮されたテキスト上のパターン照合 －データ圧縮とパターン照合の新展開－, 情報処理学会誌 43(7):763-769, 2002.
3. 竹田正幸, 福田智子: 古典和歌からの知識発見 －モビルスーツを着た国文学者－, 情報処理学会誌 43(9):941-949, 2002.

解説記事（一般誌）
1. 竹田正幸, 福田智子: 類似歌を探せ： デジタル国文学の新展開, 日経サイエンス 2002 年５月号, 2002.
2. 竹田正幸: コンピュータで和歌を読み解く, 九大広報 27 号, 2003.

新聞報道
1. 紫式部と清少納言の意外な因縁をコンピューターが発見, 2001 年 5 月 26 日朝日新聞夕刊一面トップ.
2. デジタル国文学, 2001 年 9 月 8 日 日本経済新聞朝刊文化面(裏一面).
3. 計算機にヒラメキを, 2002 年 4 月 7 日 日本経済新聞朝刊科学面.

4. 発見支援プログラム, 2003 年 11 月 12 日 朝日新聞朝刊 Labo ラボ探偵団.

招待・依頼講演

1. 竹田正幸： 文学作品からのテキストマイニング－文学における発見を支援する－, 電子情報通信学会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 (依頼講演).
2. Masayuki Takeda: Text Mining in Literary Works, PNC Annual Conference and Joint Meetings 2002. (依頼講演).
3. 竹田正幸: 文学作品におけるデータマイニング, 2002 年度統計関連学会連合大会 (招待講演).
4. 竹田正幸: 系列データからの知識発見～文学・音楽・分子生物学への応用をめぐって～, 情報処理学会九州支部 「火の国シンポジウム」, 2003. (招待講演).