

(DEWS2003 ミニサーベイ)

半構造データマイニングに関する研究動向

浅井 達哉

九州大学 大学院システム情報科学府

〒 812-8581 福岡市東区箱崎 6-10-1

t-asai@i.kyushu-u.ac.jp

高速なネットワークと安価な大容量記憶装置の発達によって、ウェブページや XML 文書に代表される半構造データ [2] がネットワーク上に蓄積されており、大規模半構造データを対象としたデータマイニング (半構造データマイニング) に対する要求が高まっている。

しかし、半構造データは関係データベースとは違い、明示的な構造をもたないので、関係データベースを対象とした従来のデータマイニング手法 [3] を、半構造データに直接適用することは困難である。そのため、木構造やグラフ構造に対する高速なマイニング手法の開発が急務となっている。

半構造データマイニングの研究は、Holder らの部分構造発見アルゴリズム Subdue [8] や Nestrov らのスキーマ抽出手法 [13]、Wang らの頻出経路列発見手法 [17] など、1990 年代なかばから顕在化し、その後、2000 年頃から、頻出部分構造を発見する問題を中心に、急激に研究が増えている [1, 4, 5, 6, 9, 10, 15, 18, 19]。

半構造データマイニングの応用は、スキーマ発見 [13, 17]、情報抽出 [14]、文書分類 [17]、巨大文書の概観 [17]、XPath を用いた検索の効率化、XML 文書の圧縮など多岐にわたる。さらに、半構造データマイニング手法は、化学物質 [7, 9, 11]、遺伝子ネットワーク [16]、インターネットのリンク構造、ネットワークログ [19]、電子回路といった複雑なデータからのデータマイニングに活用されていくことが期待されている。

本ミニサーベイでは、我々のグループで研究を進めている頻出部分構造発見アルゴリズム FREQT と OPTT、StreamT を中心に、下記に示される効率のよい半構造データマイニング手法と、その応用事例を紹介する。

- Asai らによる、順序木集合からの頻出順序木パターン発見アルゴリズム FREQT [1, 4, 5]
- Zaki による、順序木集合からの頻出順序木パターン発見アルゴリズム TreeMiner [19]
- Cong らによる、順序木集合からのラベル変数つき頻出経路列発見アルゴリズム [6]
- Inokuchi らによる、グラフ集合からの頻出部分グラフ発見アルゴリズム AGM [9]
- Kuramochi と Karypis による、グラフ集合からの頻出部分グラフ発見アルゴリズム FSG [10]
- Vanetik らによる、グラフ集合からの頻出部分グラフ発見アルゴリズム [15]
- Yan と Han による、グラフ集合からの頻出部分グラフ発見アルゴリズム gSpan [18]

参考文献

- [1] K. Abe, S. Kawasoe, T. Asai, H. Arimura, and S. Arikawa. Optimized Substructure Discovery for Semi-structured Data, In *Proc. the 6th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'02)*, 1–14, LNAI 2431, 2002.
- [2] S. Abiteboul, P. Buneman, D. Suciu, *Data on the Web*, Morgan Kaufmann, 2000.
- [3] R. Agrawal, R. Srikant, Fast Algorithms for Mining Association Rules, In *Proc. the 20th VLDB*, 487–499, 1994.
- [4] T. Asai, K. Abe, S. Kawasoe, H. Arimura, H. Sakamoto, S. Arikawa, Efficient Substructure Discovery from Large Semi-structured Data, In *Proc. the 2nd SIAM Int'l Conf. on Data Mining (SDM2002)*, 158–174, 2002.
- [5] T. Asai, H. Arimura, K. Abe, S. Kawasoe, S. Arikawa. Online Algorithms for Mining Semi-structured Data Stream, In *Proc. the 2002 IEEE Int'l Conf. on Data Mining (ICDM'02)*, 27–34, 2002.
- [6] G. Cong, L. Yi, B. Liu, K. Wang, Discovering Frequent Substructures from Hierarchical Semi-structured Data, In *Proc. SDM2002*, SIAM, 2002.
- [7] L. Dehaspe, H. Toivonen, R. D. King, Finding Frequent Substructures in Chemical Compounds, In *Proc. KDD-98*, 30–36, 1998.
- [8] L. B. Holder, D. J. Cook, S. Djoko, Substructure Discovery in the SUBDUE System, In *Proc. KDD'94*, 169–180, 1994.
- [9] A. Inokuchi, T. Washio, H. Motoda, An Apriori-Based Algorithm for Mining Frequent Substructures from Graph Data, In *Proc. PKDD 2000*, 13–23, LNAI, Springer, 2000.
- [10] M. Kuramochi, G. Karypis, Frequent Subgraph Discovery, In *Proc. IEEE ICDM'01*, 2001.
- [11] T. Matsuda, T. Horiuchi, H. Motoda, T. Washio, K. Kumazawa, N. Arai, Graph-Based Induction for General Graph Structured Data, In *Proc. DS'99*, 340–342, 1999.
- [12] T. Miyahara, Y. Suzuki, T. Shoudai, T. Uchida, K. Takahashi, H. Ueda, Discovery of Frequent Tag Tree Patterns in Semistructured Web Documents, In *Proc. PAKDD-2002*, 341–355, 2002.
- [13] S. Nestrov, S. Abiteboul, R. Motwani, Extracting Schema from Semistructured Data, In *Proc. SIGKDD'98*, 295–306, 1998.
- [14] K. Taniguchi, H. Sakamoto, H. Arimura, S. Shimozone, S. Arikawa, Mining Semi-Structured Data by Path Expressions, In *Proc. the 4th Int'l Conf. on Discovery Science (DS2001)*, LNCS 2226, 2001.
- [15] N. Vanetik, E. Gudes, E. Shimony, Computing Frequent Graph Patterns from Semistructured Data, In *Proc. IEEE ICDM'02*, 458–465, 2002.
- [16] J. T. L. Wang, B. A. Shapiro, D. Shasha, K. Zhang, C.-Y. Chang, Automated Discovery of Active Motifs in Multiple RNA Secondary Structures, In *Proc. KDD-96*, pp.70–75, 1996.
- [17] K. Wang, H. Liu, Schema Discovery for Semistructured Data, In *Proc. KDD'97*, 271–274, 1997.
- [18] X. Yan, J. Han, gSpan: Graph-Based Substructure Pattern Mining, In *Proc. IEEE ICDM'02*, 721–724, 2002.
- [19] M. J. Zaki. Efficiently Mining Frequent Trees in a Forest, In *Proc. SIGKDD 2002*, ACM, 2002.