構造化データからの知識発見に関する研究

■ 情報工学部 情報工学科 教授　正代　隆義

○ **研究分野**：知能情報学、機械学習、帰納推論
○ **キーワード**：グラフ構造データ、グラフマイニング、グラフアルゴリズム、計算論的学習

研究概要

1. 構造化データからの知識発見

ネットワーク技術の急激な進歩にともない、ウェブページに代表されるテキストデータの利用が急速に進んでいる。特に、HTMLやXMLデータ(図1)に代表される木構造データは、その規模を日増しに増大させている。また、近年では、化学化合物データ(図2)やウェブのリンク情報といったようなグラフ構造データから情報抽出を行うことに関心が集まっている。

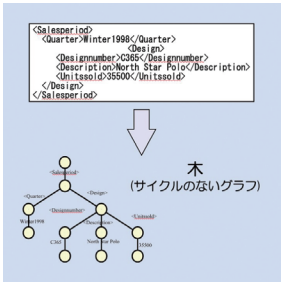

図1：XMLデータのグラフ表現 (木)

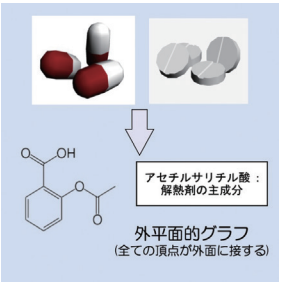

図2：化学化合物のグラフ表現 (外平面的グラフ)

本研究では、そのようなグラフ構造データを主とする大規模データベースから、多くのグラフ構造データに共通して現れる特徴的な構造を見つけ出す情報基盤技術を開発することを目的としている。

2. グラフパターンの機械学習理論と知識発見システムへの応用

グラフ構造データベースから意味のある知識を抽出するためには、まず、そのデータ中に潜むパターンをどのようにグラフパターンとして表現する(図3)か、そして、いかにしてうまくそのグラフパターンを発見するかが鍵となる。

そこで本研究では、知識発見のための理論的基盤を構築するために、計算論的学習理論における様々な学習モデルを用いて、グラフ構造データベースから現実的な時間でグラフパターン発見を行う学習アルゴリズムの開発を行う。

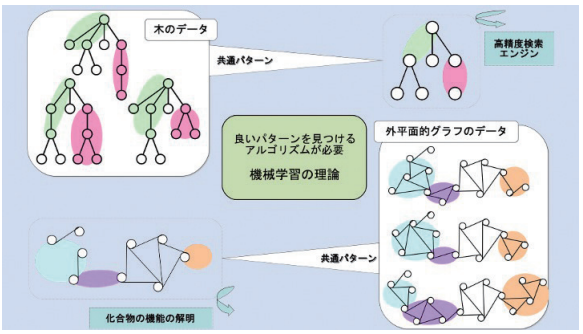

図3：グラフ構造パターンに対する高速な学習アルゴリズムの設計とグラフマイニングへの応用

利点特徴

グラフ構造をもったデータは、ウェブ、バイオ/創薬、ビジネス分析、マーケティングなど、実世界の多くの重要な場面において自然に現れる。一方で、グラフ構造を扱う計算は、組合せ爆発により「事実上の計算不可能」となることが多い。そこで、本研究では、応用場面に即した合理的な制限を入れながら、実用的な設定において学習アルゴリズムの効率化を図る。特に、本研究では、近年のAI技術の著しい進展に鑑み、ブラックボックス化した学習結果の内容理解をサポートするパターン表現と高速な学習アルゴリズムの開発を目標としている。

応用分野

本研究で開発した技術により、データが持つ内面的性質と構造的性質を結びつけるような ルールを発見する手助けとなることが期待される。これまでの研究では、HTML/XMLデータのような木構造データからの科学的知識抽出以外にも、ウェブ空間に存在する様々な構造を持つデータ、たとえば地図情報や複雑な表、化学化合物のデータベースなどから知識発見を行う学習アルゴリズムを提案している。今後も新たなグラフ構造を対象とする高速な学習アルゴリズムを開発して、それらを統合した高速なデータマイニングシステムの提供を目指している。