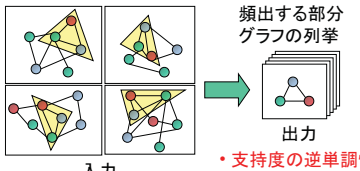
頂点により誘導される頻出グラフ系列パターンのマイニング

猪口 明博*1, *2　鷲尾 隆*1, *3
*1 大阪大学 産業科学研究所　*2 科学技術振興機構 さきがけ　*3科学技術振興機構 ERATO 湊離散構造処理系プロジェクト

背景

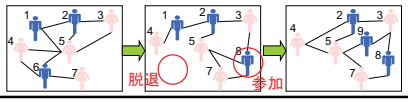
頻出グラフマイニング



- 支持度の逆単調性
- グラフ同型問題

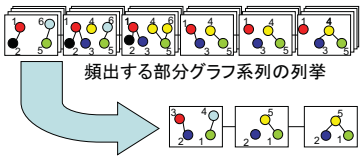
グラフ系列

- 人間関係ネットワークの変化
  - 人：頂点，人間関係：辺
- ホームページのリンク構造の変化
  - HTML文章：頂点，ハイパーリンク：辺
- 遺伝子ネットワークの変化(進化)
  - 遺伝子：頂点，相互作用：辺
- 機械の組み立て
  - 部品：頂点，隣接する部品間：辺



頻出グラフ系列マイニング

グラフ系列の集合が与えられたとき，ある頻度以上出現する頻出する部分グラフ系列の列挙すること

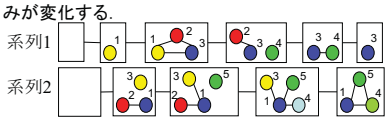


対象とするグラフ系列
•頂点数，辺数が増減する．
•頂点ラベル，辺ラベルが変化する．
•各頂点は，IDをもつ．

GTRACE

基本アイデア

仮定
グラフ系列中の連続する2つのグラフの間では，構造が大きく変化することはなく，ごく一部の構造のみが変化する．

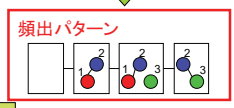


コンパイル

$\langle(vi),(vi,vi,ei,e,e),(vi,ed,ed,vd),(ei,ed,vd),(ed,vd)\rangle$
$\langle(vi,vi,vi,ei),(vi,e),(vi,ei,ei,ed,vd),(ei,ed,vd)\rangle$

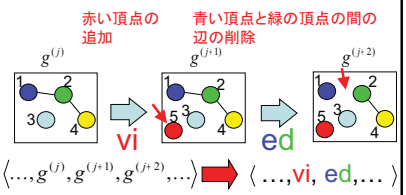
系列パターンマイニング

頻出パターンFTS (頻出変換部分系列)
$\langle(vi,vi,ei),vi,(ei,ed,vd)\rangle$



変換規則

頂点や辺の追加，削除，ラベル変更



$\langle \ldots, g^{(j)}, g^{(j+1)}, g^{(j+2)}, \ldots \rangle \Rightarrow \langle \ldots, vi, ed, \ldots \rangle$

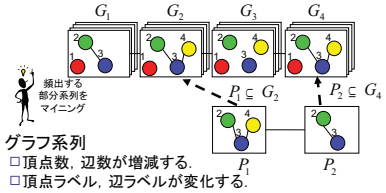
課題

GTRACEは観測されたグラフ系列中の連続する2つのグラフで，その大部分は変化せず，ごく一部の構造が変化することを仮定．

観測されたグラフ系列中の連続する2つのグラフが大きく変化する場合には，変換規則の系列が長くなり，膨大な計算時間を要する．

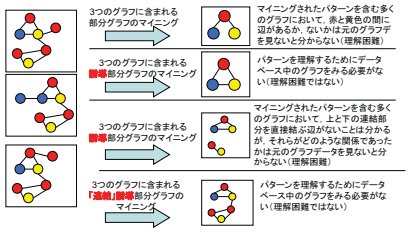
提案手法:FRISSMiner

対象とするグラフ系列



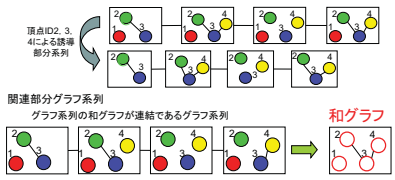
- グラフ系列
  - 頂点数，辺数が増減する．
  - 頂点ラベル，辺ラベルが変化する．
  - 各頂点は，IDをもつ．
  - グラフ系列中の連続する2つのグラフの間で，**構造が大きく変化する．**
  - 個々のグラフが大きく，系列が長い．

理解容易な頻出部分グラフとは？



FRISS

FRISS (Frequent, Relevant, and Induced Subgraph Subsequence)



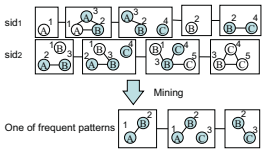
問題定義

グラフ系列集合 $DB = \{d \mid d = \langle g^{(1)} g^{(2)} \cdots g^{(n)} \rangle\}$ と閾値(最小支持度)σ'が入力として与えられたとき，DB中の頻出するグラフ系列パターンを全て列挙すること

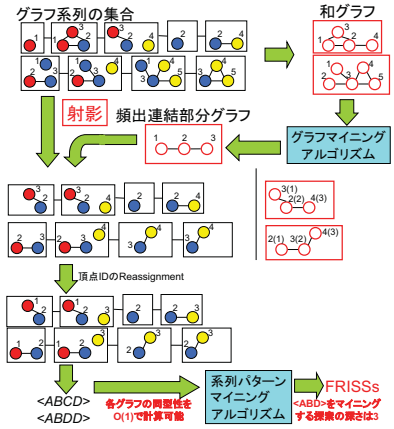
ただし，

1.パターンは元のグラフに誘導部分グラフ系列として含まれているものとする．

2.パターンの和グラフは連結であるとする．



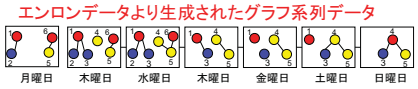
アルゴリズム

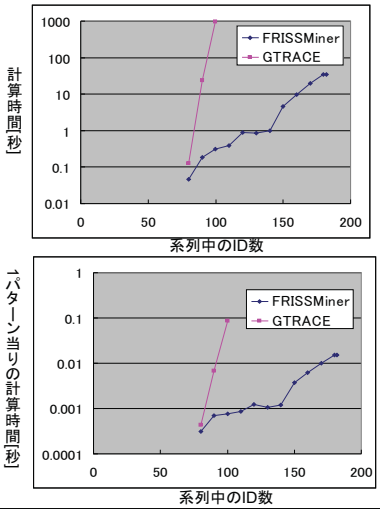


評価実験

エンロンデータ
　電子メールの履歴データ
　123週
前処理
　頂点ID: E-mailアドレス (人)
　辺: ある日にコミュニケーションをとった2名
　頂点ラベル: 8種のラベルのいずれか
　　CEO, Employee, Director, Manager,
　　Lawyer, President, Trader, Vice President

エンロンデータより生成されたグラフ系列データ





- グラフ系列中のID数の増加に対して，計算時間は指数関数的に増加(上図)
- グラフ系列中のID数の増加に対して，1パターン当りの計算時間は指数関数的に増加(下図)
- 提案手法FRISSMinerは，グラフ系列が長く，グラフ系列の各グラフが大きいグラフにも適用可能

まとめ

マイニングするパターンを頻出関連誘導部分グラフ系列(FRISS)に制限することによる効果
- FRISSは理解困難ではない．
  - FRISSを理解するためにデータベース中のグラフ系列をみる必要がない．
- 誘導部分グラフ系列に限定することで，パターンの候補が減るため，計算時間は短くなる．
- 関連部分グラフ系列に限定することで，パターンの候補が減るため，計算時間は短くなる．
- 射影後のグラフ系列をアイテムの系列で，表現できるので，効率良くパターンをマイニングできる．