



HOKKAIDO UNIVERSITY

Title	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
Author(s)	猪口, 明博; 鷲尾, 隆
Description	ERATO 湊離散構造処理系プロジェクトシンポジウム (第1回) : 第9回情報科学技術フォーラム(FIT2010) イベント企画セッション. 2010年9月8日 (水). 九州大学伊都キャンパス.
Relation	2010年度科学技術振興機構ERATO湊離散構造処理系プロジェクト講究録. p.373.
Issue Date	2011-06
Doc URL	https://hdl.handle.net/2115/48370
Type	conference presentation
File Information	02.FIT_inokuchi.pdf



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

頻出する部分グラフの列挙

出力

- 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

頻出する部分グラフ系列の列挙

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い, ごく一部の構造のみが変化する.

系列1:

系列2:

コンパイル

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

系列パターンマイニング

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

頻出パターン

変換規則

頂点や辺の追加, 削除, ラベル変更

赤い頂点の追加

青い頂点と緑の頂点の間の辺の削除

$g^{(j)}$ → $g^{(j+1)}$ → $g^{(j+2)}$

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

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

G_1, G_2, G_3, G_4

頻出する部分系列をマイニング

$P_1 \subseteq G_2, P_2 \subseteq G_4$

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?

3つのグラフに含まれる部分グラフのマイニング

3つのグラフに含まれる部分グラフのマイニング

3つのグラフに含まれる部分グラフのマイニング

マイニングされたパターンを含む多くのグラフにおいて, 赤と黄色の間に辺があるか, ないかは元のグラフで見ないと分からない(理解困難)

パターンを理解するためにデータベース中のグラフをみる必要がない(理解困難ではない)

マイニングされたパターンを含む多くのグラフにおいて, 上と下の連結部分を直接見ることがない. これは分かるが, それ以外の部分で繋がったものは元のグラフデータを見ないと分からない(理解困難)

パターンを理解するためにデータベース中のグラフをみる必要がない(理解困難ではない)

FRISS

FRISS (Frequent, Relevant, and Induced Subgraph Subsequence)

誘導部分グラフ系列

頂点ID: 3, 4による誘導部分系列

関連部分グラフ系列

グラフ系列の和グラフが連結であるグラフ系列

和グラフ

問題定義

グラフ系列集合 $DB = \{d | d = \langle g^{(1)} g^{(2)} \dots g^{(n)} \rangle\}$ 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,

- 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.
- パターンの和グラフは連結であるとする.

射影 頻出連結部分グラフ

和グラフ

グラフ系列の集合

射影

頻出連結部分グラフ

グラフマイニングアルゴリズム

頻出パターン

各グラフの関連性を $O(1)$ で計算可能

系列パターンマイニングアルゴリズム

FRISSs

$\langle ABCD \rangle$

$\langle ABD \rangle$

$\langle 2(1), 3(2), 4(3) \rangle$

$\langle 2(1), 3(2) \rangle$

$\langle 4(3) \rangle$

各グラフの関連性を $O(1)$ で計算可能

系列パターンマイニングアルゴリズム

FRISSs

$\langle ABD \rangle$ をマイニングする探索の置きは?

評価実験

エンロンデータ

電子メールの履歴データ
 123週

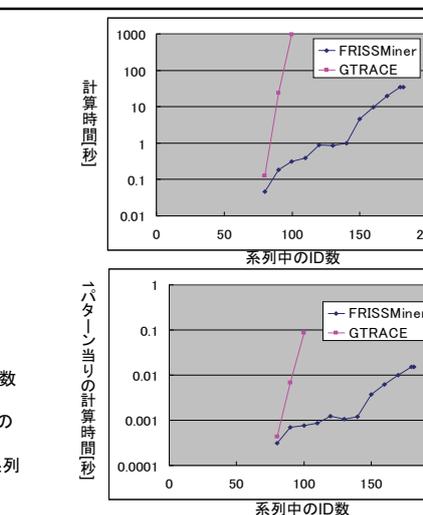
前処理

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

月曜日 木曜日 水曜日 木曜日 金曜日 土曜日 日曜日

- 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.