Foundations of Query Processing

R19-4: Dynamic Programming: The Next Step

Takeshi Yamamuro

20150516@ICDE2015study

Outline

- 結合演算（join）と集約演算（group-by）で構成される関係代数式の等価性を包括的に定義

  →より効率的な演算順序を探索することが可能に

- 新たな等価性で拡大した探索空間に対して効率的な演算順序の選択方法を議論

  →実験から|R|が7までは全列挙が可能，10までは探索の枝刈り，それ以上は動的計画法（DP）を前提とした経験的なアプローチを適用することが妥当

Backgrounds

- 結合演算→集約演算は頻出パタン
  - 正規化された関係Rを結合後に集計処理する場合など

```
SELECT
   a.gid, SUM(b.value), AVG(b.value)
FROM
   a LEFT OUTER JOIN b ON (a.id= b.id)
GROUP BY a.gid
```

- 結合前に集約することで効率化の可能性
  - 結合選択率（join selectivity）と組（tuple）の重複度が高い
  - Hyper [9]において100x～1000xの改善を確認

Backgrounds

- 90年前半に内部結合演算と集約演算のみに関する等価性に関する提案 [4][5][6][7][8]

- 本論文では外部結合を含めた5つの結合演算に関する包括的な等価性を議論
  - left semijoin
  - left antijoin
  - left outerjoin
  - full outerjoin
  - groupjoin [11]

ex.) 外部結合（outerjoin）の等価性

- 外部結合の左関係e_1に集約演算をpush-down

集約演算Γ → 集約する属性Gと集約関数F（SUMやAVGなど）

$$\Gamma_{G;F}(e_1 \mathbin{⟗}_q e_2) \equiv$$

関係e_1とe_2を条件qでfull outerjoin

$$\Gamma_{G;(F_2 \otimes c_1) \circ F_1^2}(\Gamma_{G_1^+;F_1^1 \circ (c_1:\mathrm{count}(*))}(e_1) \mathbin{⟗}_q^{F_1^1(\{\bot\}),c_1:1;-} e_2)$$

左関係e_1にpush-down

ex.) 外部結合（outerjoin）の等価性

- 外部結合の左関係e_1に集約演算をpush-down

$$\Gamma_{G;F}(e_1 \mathbin{⟗}_q e_2) \equiv$$

集約関数F*のdecomposability/splittabilityを考慮して分割

$$\Gamma_{G;(F_2 \otimes c_1) \circ F_1^2}(\Gamma_{G_1^+;F_1^1 \circ (c_1:\mathrm{count}(*))}(e_1) \mathbin{⟗}_q^{F_1^1(\{\perp\}),c_1:1;-} e_2)$$

- G^+_1はe_1の集約属性G_1と結合属性J_1の和集合
- G^+_1と分割された$F^1_1 \circ (c_1:\mathrm{count}(*))$を用いて集約処理

* 集約関数Fのdecomposability/splittabilityの定義は論文内Ⅱ-A. 2)/3)を参照

ex.) 外部結合（outerjoin）の等価性

- 外部結合の左関係e_1に集約演算をpush-down

$$\Gamma_{G;F}(e_1 \mathbin{⟗}_q e_2) \quad \equiv$$

$$\Gamma_{G;(F_2 \otimes c_1) \circ F_1^2}(\Gamma_{G_1^+;F_1^1 \circ (c_1:\mathrm{count}(*))}(e_1) \mathbin{⟗}_q^{F_1^1(\{\perp\}),c_1:1;-} e_2)$$

外部結合の結果に対して再度集約処理を実施*

* 最上位の集約演算の省略条件はⅢ-B（Eliminating the Top Grouping）を参照

包括的な等価性を定義: 式(10)～(41)

Eager/Lazy Groupby-Count

$$\Gamma_{G;F}(e_1 \bowtie_q e_2) \equiv \Gamma_{G;(F_2\otimes c_1)\circ F_1^2}(\Gamma_{G_1^+;F_1^1\circ(c_1:\mathrm{count}(*))}(e_1) \bowtie_q e_2) \quad (10)$$

$$\Gamma_{G;F}(e_1 \bowtie_q e_2) \equiv \Gamma_{G;(F_2\otimes c_1)\circ F_1^2}(\Gamma_{G_1^+;F_1^1\circ(c_1:\mathrm{count}(*))}(e_1) \bowtie_q e_2) \quad (11)$$

$$\Gamma_{G;F}(e_1 \bowtie_q e_2) \equiv \Gamma_{G;(F_2\otimes c_1)\circ F_1^2}(\Gamma_{G_1^+;F_1^1\circ(c_1:\mathrm{count}(*))}(e_1) \bowtie_q^{F_1^1(\{\perp\}),c_1:1;-} e_2) \quad (12)$$

$$\Gamma_{G;F}(e_1 \bowtie_q e_2) \equiv \Gamma_{G;(F_1\otimes c_2)\circ F_2^2}(e_1 \bowtie_q \Gamma_{G_2^+;F_2^1\circ(c_2:count(*))}(e_2)) \quad (13)$$

$$\Gamma_{G;F}(e_1 \bowtie_q e_2) \equiv \Gamma_{G;(F_1\otimes c_2)\circ F_2^2}(e_1 \bowtie_q^{F_2^1(\{\perp\}),c_2:1} \Gamma_{G_2^+;F_2^1\circ(c_2:\mathrm{count}(*))}(e_2)) \quad (14)$$

$$\Gamma_{G;F}(e_1 \bowtie_q e_2) \equiv \Gamma_{G;(F_1\otimes c_2)\circ F_2^2}(e_1 \bowtie_q^{-;F_2^1(\{\perp\}),c_2:1} \Gamma_{G_2^+;F_2^1\circ(c_2:\mathrm{count}(*))}(e_2)) \quad (15)$$

Eager/Lazy Group-by

$$\Gamma_{G;F}(e_1 \bowtie_q e_2) \equiv \Gamma_{G;F_1^2}(\Gamma_{G_1^+;F_1^1}(e_1) \bowtie_q e_2) \quad (16)$$

$$\Gamma_{G;F}(e_1 \bowtie_q e_2) \equiv \Gamma_{G;F_1^2}(\Gamma_{G_1^+;F_1^1}(e_1) \bowtie_q e_2) \quad (17)$$

$$\Gamma_{G;F}(e_1 \bowtie_q e_2) \equiv \Gamma_{G;F_1^2}(\Gamma_{G_1^+;F_1^1}(e_1) \bowtie_q^{F_1^1(\{\perp\});-} e_2) \quad (18)$$

$$\Gamma_{G;F}(e_1 \bowtie_q e_2) \equiv \Gamma_{G;F_2^2}(e_1 \bowtie_q \Gamma_{G_2^+;F_2^1}(e_2)) \quad (19)$$

$$\Gamma_{G;F}(e_1 \bowtie_q e_2) \equiv \Gamma_{G;F_2^2}(e_1 \bowtie_q^{F_2^1(\{\perp\})} \Gamma_{G_2^+;F_2^1}(e_2)) \quad (20)$$

$$\Gamma_{G;F}(e_1 \bowtie_q e_2) \equiv \Gamma_{G;F_2^2}(e_1 \bowtie_q^{-;F_2^1(\{\perp\})} \Gamma_{G_2^+;F_2^1}(e_2))$$

Eager/Lazy Count

$$\Gamma_{G;F}(e_1 \bowtie_q e_2) \equiv \Gamma_{G;(F_2\otimes c_1)}(\Gamma_{G^+;$$

$$\Gamma_{G;F}(e_1 \bowtie_q e_2) \equiv$$

(25)

(26)

$$\Gamma_{G;F}(e_1 \bowtie_q e_2) \equiv \Gamma_{G;F_2^2}(e_1 \bowtie_q \Gamma_{G_2^+;(c_2:\mathrm{count}(*))}(e_2)) \quad (27)$$

Eager/Lazy

$$\Gamma_{G;F}(e_1 \bowtie_q e_2) \equiv \Gamma_{G;(F_1^2\otimes c_2)}(\Gamma_{G_1^+;F_1^1}(e_1) \bowtie_q \Gamma_{G_2^+;c_2:\mathrm{count}(*)}(e_2)) \quad (28)$$

$$\Gamma_{G;F}(e_1 \bowtie_q e_2) \equiv \Gamma_{G;(F_1^2\otimes c_2)}(\Gamma_{G_1^+;F_1^1}(e_1) \bowtie_q^{c_2:1} \Gamma_{G_2^+;c_2:\mathrm{count}(*)}(e_2)) \quad (29)$$

$$\Gamma_{G;F}(e_1 \bowtie_q e_2) \equiv \Gamma_{G;(F_1^2\otimes c_2)}(\Gamma_{G_1^+;F_1^1}(e_1) \bowtie_q^{F_1^1(\{\perp\});c_2:1} \Gamma_{G_2^+;(c_2:\mathrm{count}(*))}(e_2)) \quad (30)$$

$$\Gamma_{G;F}(e_1 \bowtie_q e_2) \equiv \Gamma_{G;(F_2^2\otimes c_1)}(\Gamma_{G_1^+;c_1:\mathrm{count}(*)}(e_1) \bowtie_q \Gamma_{G_2^+;F_2^1}(e_2)) \quad (31)$$

$$\Gamma_{G;F}(e_1 \bowtie_q e_2) \equiv \Gamma_{G;(F_2^2\otimes c_1)}(\Gamma_{G_1^+;c_1:\mathrm{count}(*)}(e_1) \bowtie_q^{F_2^1(\{\perp\})} \Gamma_{G_2^+;F_2^1}(e_2)) \quad (32)$$

$$\Gamma_{G;F}(e_1 \bowtie_q e_2) \equiv \Gamma_{G;(F_2^2\otimes c_1)}(\Gamma_{G_1^+;(c_1:\mathrm{count}(*))}(e_1) \bowtie_q^{c_1:1;F_2^1(\{\perp\})} \Gamma_{G_2^+;F_2^1}(e_2)) \quad (33)$$

Eager/Lazy Split (with $\Gamma^2 := \Gamma_{G;(F_1^2\otimes c_2)\circ(F_2^2\otimes c_1)}$):

$$\Gamma_{G;F}(e_1 \bowtie_q e_2) \equiv \Gamma_{G;(F_1^2\otimes c_2)\circ(F_2^2\otimes c_1)}(\Gamma_{G_1^+;F_1^1\circ(c_1:\mathrm{count}(*))}(e_1) \bowtie_q \Gamma_{G_2^+;F_2^1\circ(c_2:\mathrm{count}(*))}(e_2)) \quad (34)$$

$$\Gamma_{G;F}(e_1 \bowtie_q e_2) \equiv \Gamma_{G;(F_1^2\otimes c_2)\circ(F_2^2\otimes c_1)}(\Gamma_{G_1^+;F_1^1\circ(c_1:\mathrm{count}(*))}(e_1) \bowtie_q^{F_2^1(\{\perp\}),c_2:1} \Gamma_{G_2^+;F_2^1\circ(c_2:\mathrm{count}(*))}(e_2)) \quad (35)$$

$$\Gamma_{G;F}(e_1 \bowtie_q e_2) \equiv \Gamma^2(\Gamma_{G_1^+;F_1^{1,1}\circ(c_1:\mathrm{count}(*))}(e_1) \bowtie_q^{F_1^{1,1}(\{\perp\}),c_1:1;F_2^{1,1}(\{\perp\}),c_2:1} \Gamma_{G_2^+;F_2^{1,1}\circ(c_2:\mathrm{count}(*))}(e_2)) \quad (36)$$

Others

$$\Gamma_{G;F}(e_1 \ltimes_q e_2) \equiv \Gamma_{G;F}(e_1) \ltimes_q e_2 \quad (\mathcal{F}(q) \cap \mathcal{A}(e_1)) \subseteq G \quad (37)$$

$$\Gamma_{G;F}(e_1 \triangleright_q e_2) \equiv \Gamma_{G;F}(e_1) \triangleright_q e_2 \quad (\mathcal{F}(q) \cap \mathcal{A}(e_1)) \subseteq G \quad (38)$$

$$\Gamma_{G;F}(e_1 \bowtie_{J_1\theta J_2;\overline{F}} e_2) \equiv \Gamma_{G;(F_2\otimes c_1)\circ F_1^2}(\Gamma_{G_1^+;F_1^1\circ(c_1:\mathrm{count}(*))}(e_1) \bowtie_{J_1\theta J_2;\overline{F}} e_2)). \quad (39)$$

$$\Gamma_{G;F}(e_1 \bowtie_{J_1\theta J_2;\overline{F}} e_2) \equiv \Gamma_{G;F_1^2}(\Gamma_{G_1^+;F_1^1}(e_1) \bowtie_{J_1\theta J_2;\overline{F}} e_2)) \quad (40)$$

$$\Gamma_{G;F}(e_1 \bowtie_{J_1\theta J_2;\overline{F}} e_2) \equiv \Gamma_{G;(F_2\otimes c_1)}(\Gamma_{G_1^+;(c_1:\mathrm{count}(*))}(e_1) \bowtie_{J_1\theta J_2;\overline{F}} e_2)) \quad (41)$$

Not Explained

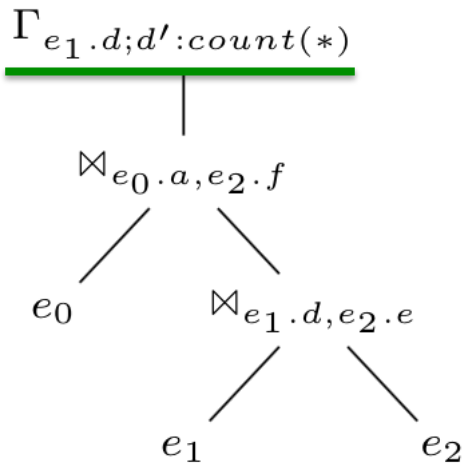
演算順序最適化と動的計画法

- 演算順序の解空間$O(2^{2|R|-1}\#ccp)$に対して動的計画法を用いて最適解を探索
  - ccp(**c**sg-**c**mp-**p**air)*はjoin graphを2つに分割して構成される部分集合S_1とS_2
  - PostgreSQLの最適化においても$|R|<11$の条件では動的計画法を適用

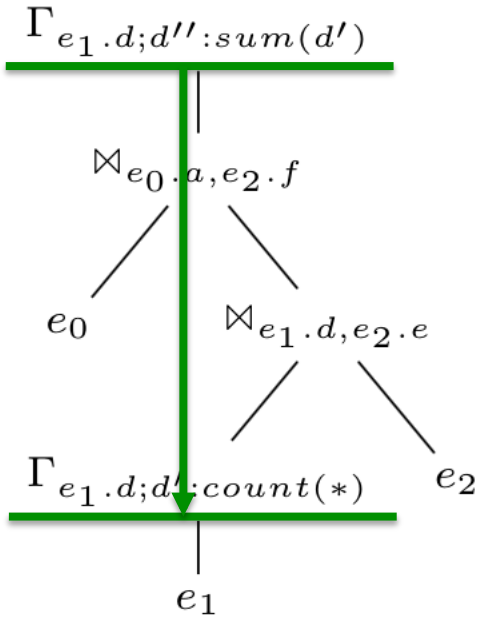
* ccpの正確な定義はⅣ-BのDefinition3を参照

動的計画法と部分問題独立性

- 結合と集約の順序問題は互いの部分問題が依存
  - 最適性原理（Bellman’s Principle of Optimality）が成り立たないため，一般的なDPで最適解が得られる保証はない

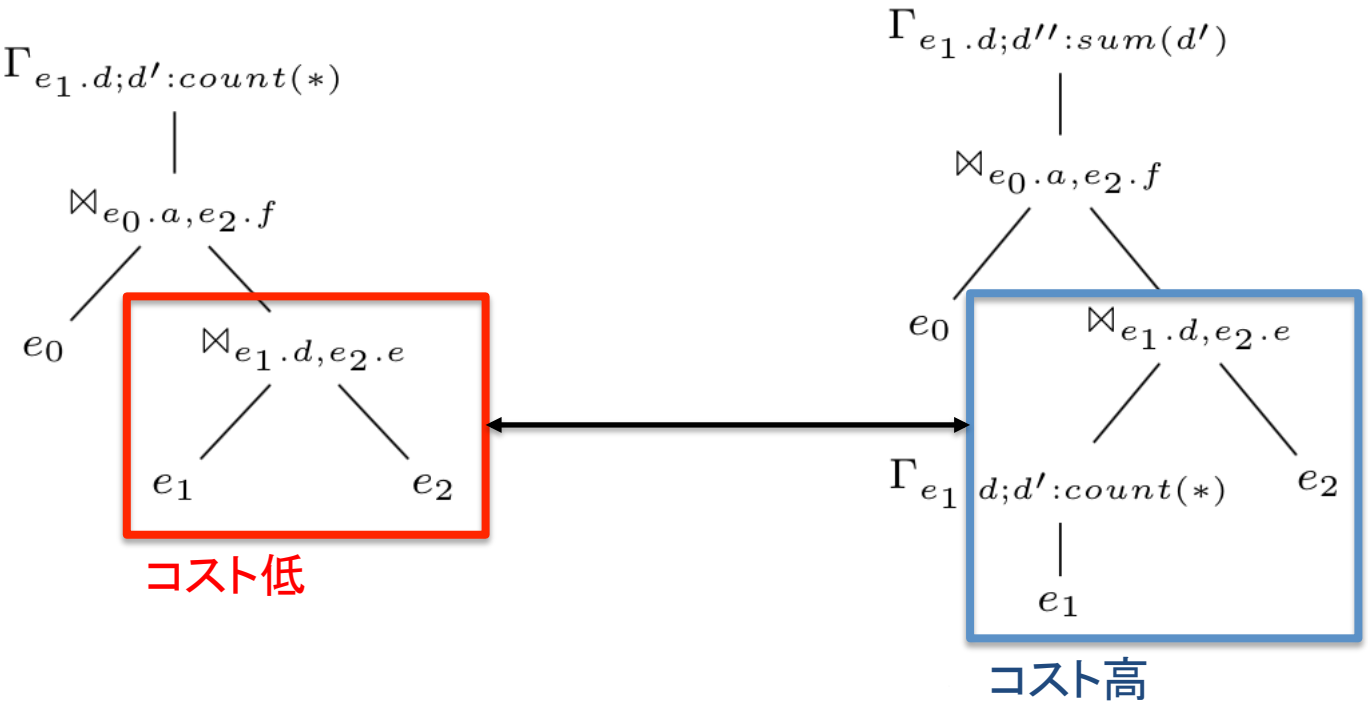


集約演算をpush-down



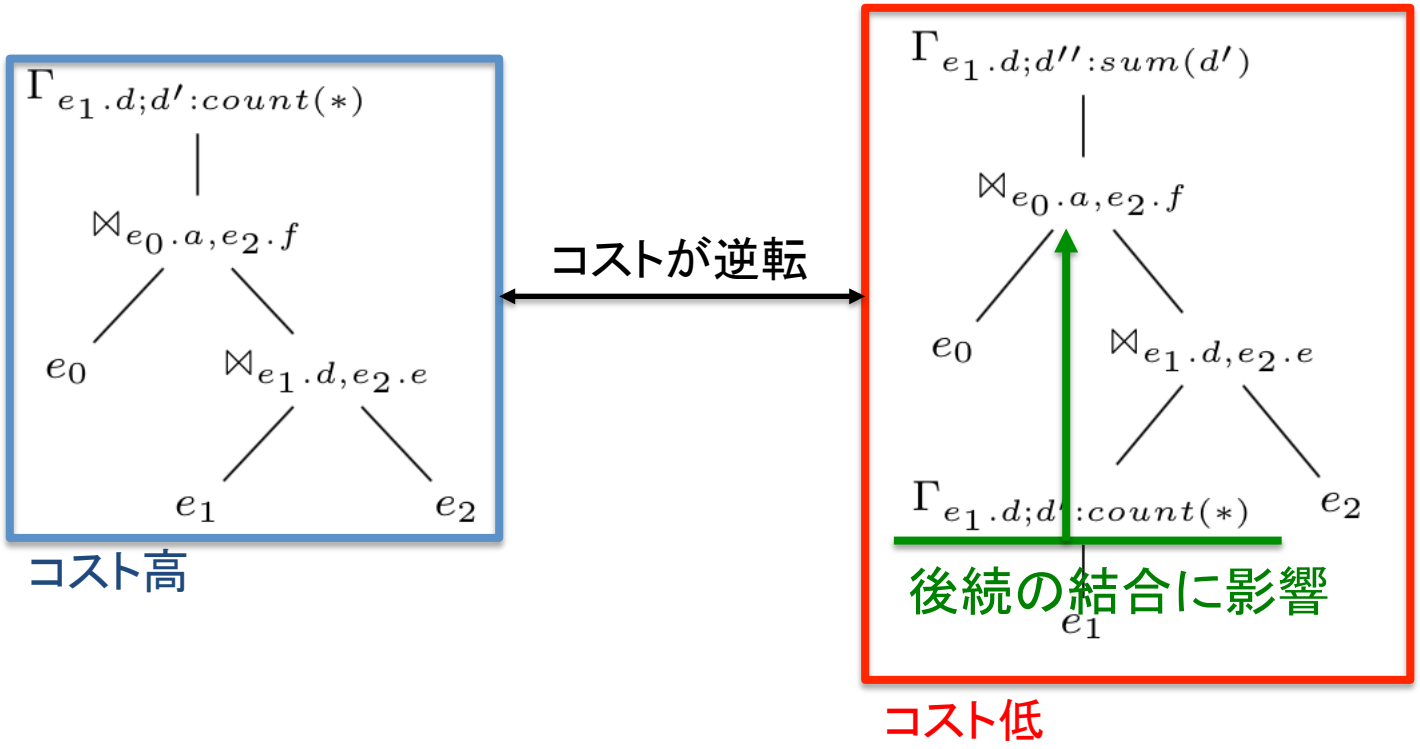
動的計画法と部分問題独立性

- 結合と集約の順序問題は互いの部分問題が依存
  - 最適性原理（Bellman's Principle of Optimality）が成り立たないため，一般的なDPで最適解が得られる保証はない



動的計画法と部分問題独立性

- 結合と集約の順序問題は互いの部分問題が依存
  - 最適性原理（Bellman's Principle of Optimality）が成り立たないため，一般的なDPで最適解が得られる保証はない



適用する4つの探索手法

- 最適解を保証するアプローチ
  - EA-All: 全列挙
  - EA-Prune: EA-All + 枝刈り
- 経験的なアプローチ
  - H1: 部分問題で最適な解を1つメモしてDPを適用
  - H2: H1 + 集約処理のpush-downの影響を考慮

実験結果

- |R|=3～20として，10,000パタンのランダムな演算順序を初期値として探索を実行

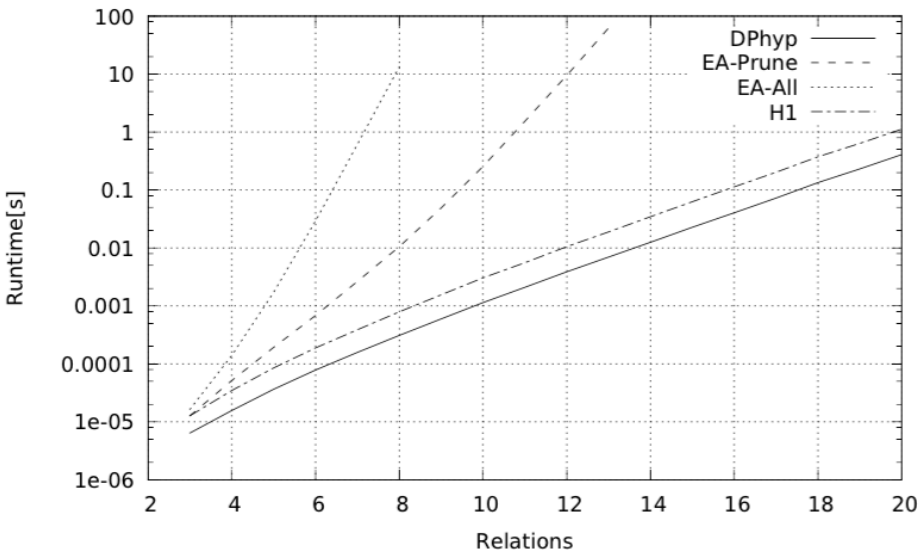


Fig.16 実行時間

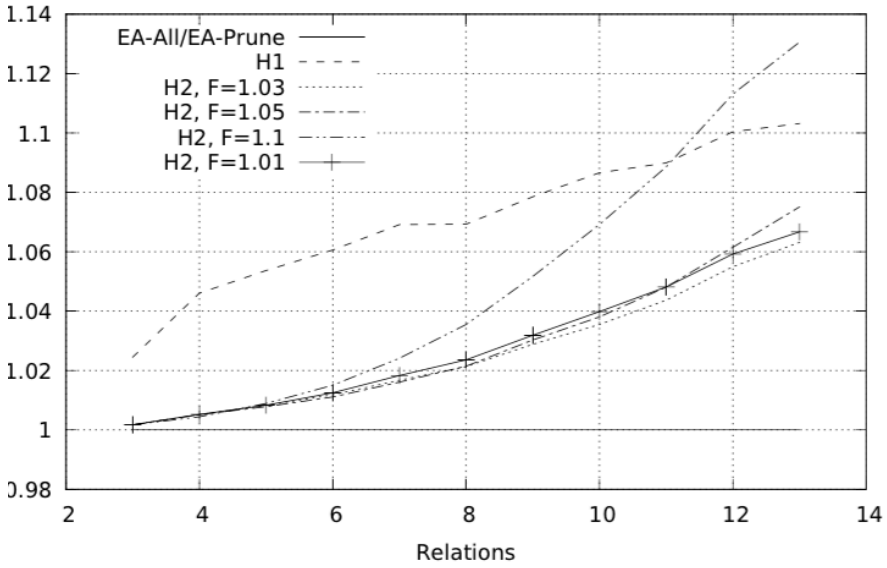


Fig.17 最適解からの乖離