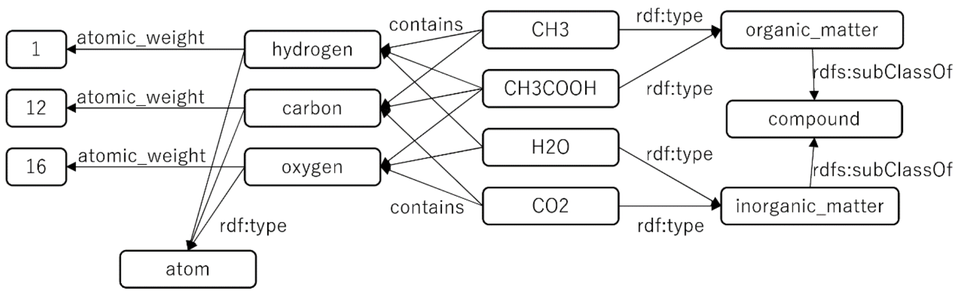
SPARQLにおける集約関数の設計とクエリ書き換え

情報・ネットワーク工学専攻 兼岩研究室　平山 健太

RDFデータ

RDFトリプル：
　主語，**述語**，**目的語**の三つ組.
RDFデータ：
　RDFトリプルの集合，グラフ構造をなす.



SPARQLクエリ：
　RDFデータに対する検索のための
　形式化された質問．構文はSQLに類似する.

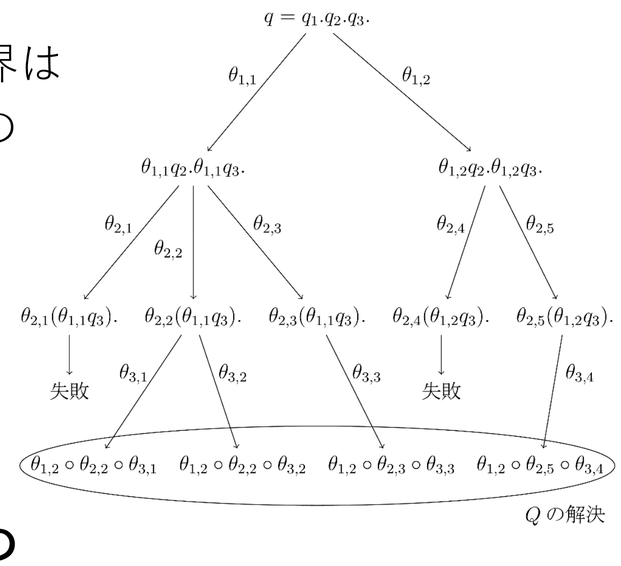
```
SELECT DISTINCT ?X
WHERE
{
   ?X rdf:type ub:GraduateStudent .
   ?X ub:memberOf ?Z .
   ?X ub:undergraduateDegreeFrom ?Y .
}
```

SPARQLクエリの処理

RDFデータの検索：
　クエリに含まれる変数(?X，?Yなど)に
　あてはまる**主語**，**述語**，**目的語**の組み合わせを
　すべて見つけること.

検索時にはクエリ内の変数にあてはまるリソースの組み合わせの探索が必要.

探索にかかる時間の上界は
クエリに含まれる変数の
個数の増加に伴って
指数関数的に増加する.



検索時間の削減のための工夫が必要.

藤原 浩司, 兼岩 憲. "大規模RDFグラフのための効率的なクエリ解決". 平成24年度第3回情報処理学会東北支部研究会資料 p. 2 図1を引用

提案手法

提案手法の適用対象：
　DISTINCTキーワードを用いる(検索結果から
　重複を排除する)クエリ

クエリに含まれる変数の**依存関係を破壊しない**よう
複数の小クエリに分割

各小クエリの検索結果を
SPARQLの集約関数で
集計し，もとのクエリの
結果を求める.

もとのクエリの変数の個数
n, 最大の小クエリの変数の
個数n', 定数aによって
計算量の上界は
$O(a^n) \rightarrow O(a^{n'})$ に改善される.

```
SELECT ?X
WHERE
{
  {
    SELECT DISTINCT ?X
    {
      ?X rdf:type ub:GraduateStudent .
    }
  }
  UNION
  {
    SELECT DISTINCT ?X
    {
      ?X ub:memberOf ?Z .
    }
  }
  UNION
  {
    SELECT DISTINCT ?X
    {
      ?X ub:undergraduateDegreeFrom ?Y .
    }
  }
}
GROUP BY (?X) HAVING (COUNT(?X) = 3)
```

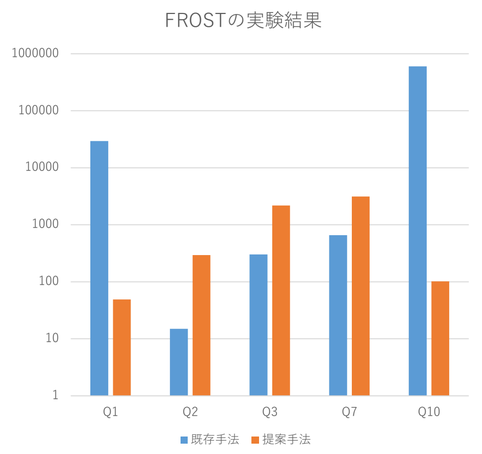
実験

複数のRDFデータベースを用いた評価実験
LUBMデータに対する10個のクエリの処理時間を計測

結果：

高速化の**効果が出たクエリ**と
出なかったクエリの両方が
存在

利用したRDFデータベースに
よっても高速化の程度が変化



今後の課題

- 高速化の効果が見込めるクエリとそうでないクエリの分類
- RDFデータについての統計情報を蓄積するデータ構造の設計