【VLDB2010勉強会】

Session 11: Ranking Queries

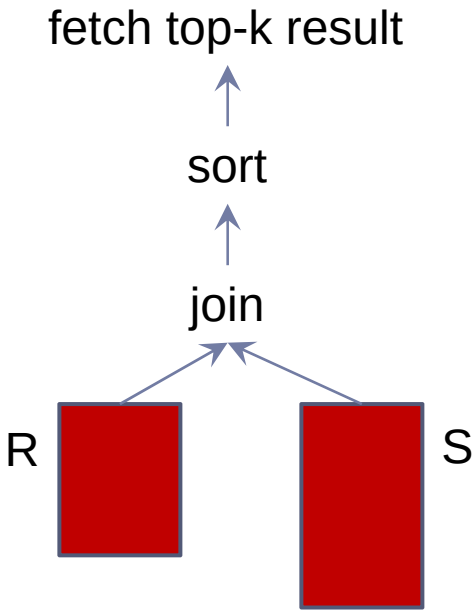
担当: 波多野 賢治（同志社大学）

Proximity Rank Join (Martinenghi et al.)

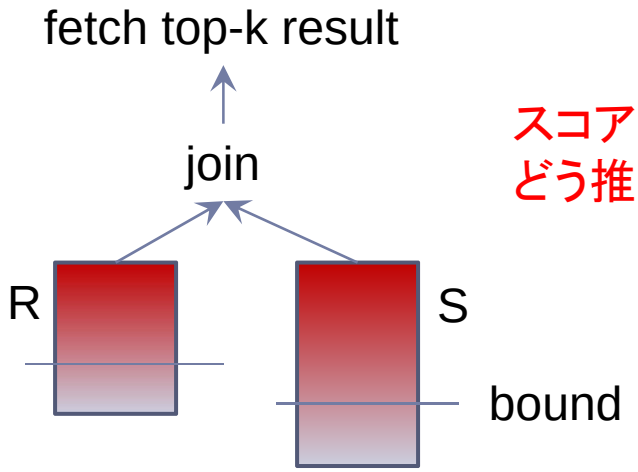
- European Research Council の Search Computing プロジェクトの成果
  - Stefano Ceri が (おそらく) プロジェクトリーダー
- Top-k Query Execution における問題解決を目指した研究
  - HRJN/HRJN* などで行われていた corner bound な方法よりも tight bound な方法が効率的
    - リレーションで扱われている値は，単なる値ではなくベクトル

Proximity Rank Join (Martinenghi et al.)

- Top-k Query Execution の問題



すべての組合せを
計算するので
処理コストがかかる



スコア上位のタップルを
どう推定するかが問題

RとSをスコア順にソートし，それぞれのリレーションから
top-k result を得るために必要なスコア上位のタップルを
推定した上で join

Proximity Rank Join (Martinenghi et al.)

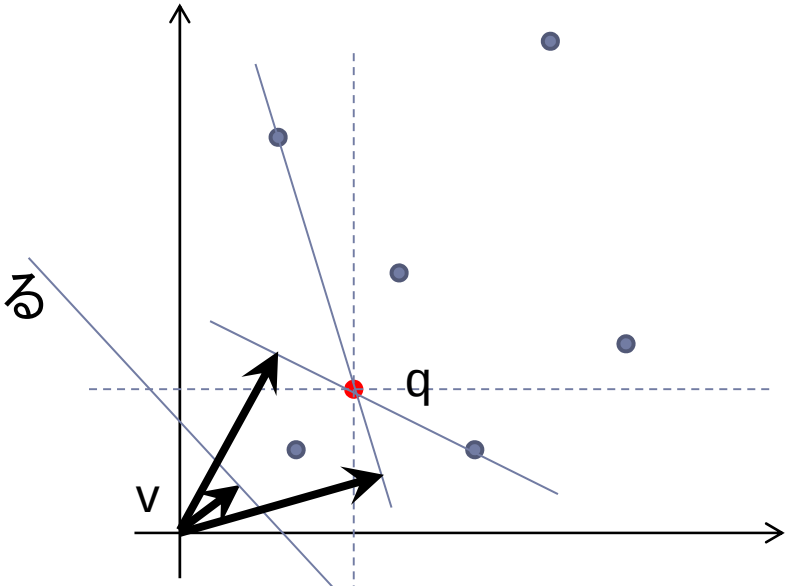
- Corner bound (HRJN etc.)
  - Join に関わるリレーション双方にアクセスしなければ，それぞれのリレーションの bound $t(\tau)$（上位のタップルと判断できるライン）を推定できない
- Tight bound (提案手法)
  - アクセスしたリレーションによって，bound (t_m) を推定できる

Table 3: Partial combinations formed with the tuples of Table 1.

M	$\tau \in PC(M)$	$t(\tau)$	t_M	
$\emptyset$	()	-19.2	-19.2	
{1}	$\tau_1^{(1)}$	-20.6	-19.2	R_1の bound
	$\tau_1^{(2)}$	-19.2		
{2}	$\tau_2^{(1)}$	-12.8	-12.8	R_2の bound
	$\tau_2^{(2)}$	-19.4		
{3}	$\tau_3^{(1)}$	-12.8	-12.8	R_3の bound
	$\tau_3^{(2)}$	-20.1		
{1,2}	$\tau_1^{(1)} \times \tau_2^{(1)}$	-16.0	-13.5	R_1,R_2 のbound
	$\tau_1^{(1)} \times \tau_2^{(2)}$	-24.0		
	$\tau_1^{(2)} \times \tau_2^{(1)}$	-13.5		
	$\tau_1^{(2)} \times \tau_2^{(2)}$	-20.4		
{1,3}	$\tau_1^{(1)} \times \tau_3^{(1)}$	-16.0	-13.5	R_1,R_3 のbound
	$\tau_1^{(1)} \times \tau_3^{(2)}$	-22.0		
	$\tau_1^{(2)} \times \tau_3^{(1)}$	-13.5		
	$\tau_1^{(2)} \times \tau_3^{(2)}$	-26.4		
{2,3}	$\tau_2^{(1)} \times \tau_3^{(1)}$	-7.0	-7.0	R_2,R_3 のbound
	$\tau_2^{(1)} \times \tau_3^{(2)}$	-21.0		
	$\tau_2^{(2)} \times \tau_3^{(1)}$	-13.1		
	$\tau_2^{(2)} \times \tau_3^{(2)}$	-26.8		

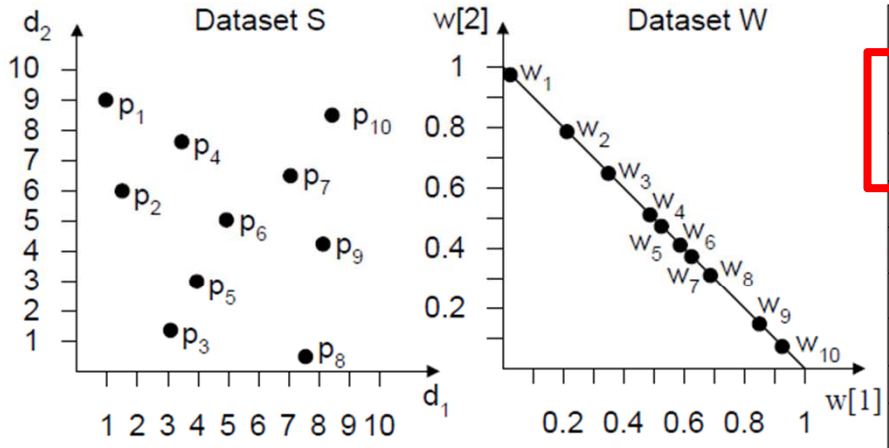
Identifying the Most Influential Data Objects with Reverse Top-k Queries (Vlachou et al.)

- Reverse Top-k Query Execution における問題解決を目指した研究
  - ICDE2010 は単にどのように処理すれば現実的なレベルで実現できるかを述べたに過ぎない
- モデル
  - Top-k Query
    - ベクトル v に対して垂線を引く
    - 垂線を原点より順に離していく
    - 各点と重なった順が Top-k となる
  - Reverse Top-k Query
    - ある点に対して座標軸をとる
    - 第2, 4象限の点とqを結ぶ
    - それぞれの直線の傾きがqの逆Top-2の解ボーダ
    - 逆Top-2の解は2直線に対する原点からの垂線ベクトル



Identifying the Most Influential Data Objects with Reverse Top-k Queries (Vlachou et al.)

- 問題設定
  - 垂線ベクトル間に含まれるオブジェクトが，逆 Top-k の候補だが，それらすべてを解の候補として調べることは非効率
  - （おそらく）すべての点を Dataset W の形で表現し，Dataset S 上の点により影響を与える (Most Influential) オブジェクトのリスト ($RTOP_2(p)$) の個数 $f_I^k()$ から精査すべきオブジェクトを選択 (Skyband-based Algorithm (SB))



	p_i	$RTOP_2(p)$	$f_I^k()$
1	p_3	$\{w_1,w_2,w_3,w_4,w_5,w_6,w_7,w_8\}$	8
2	p_2	$\{w_6,w_7,w_8,w_9,w_{10}\}$	5
3	p_8	$\{w_1,w_2,w_3\}$	3
4	p_5	$\{w_4,w_5\}$	2
5	p_1	$\{w_9,w_{10}\}$	2
6	p_4	{ }	0
		...	
10	p_{10}	{ }	0

Most Influential Objects

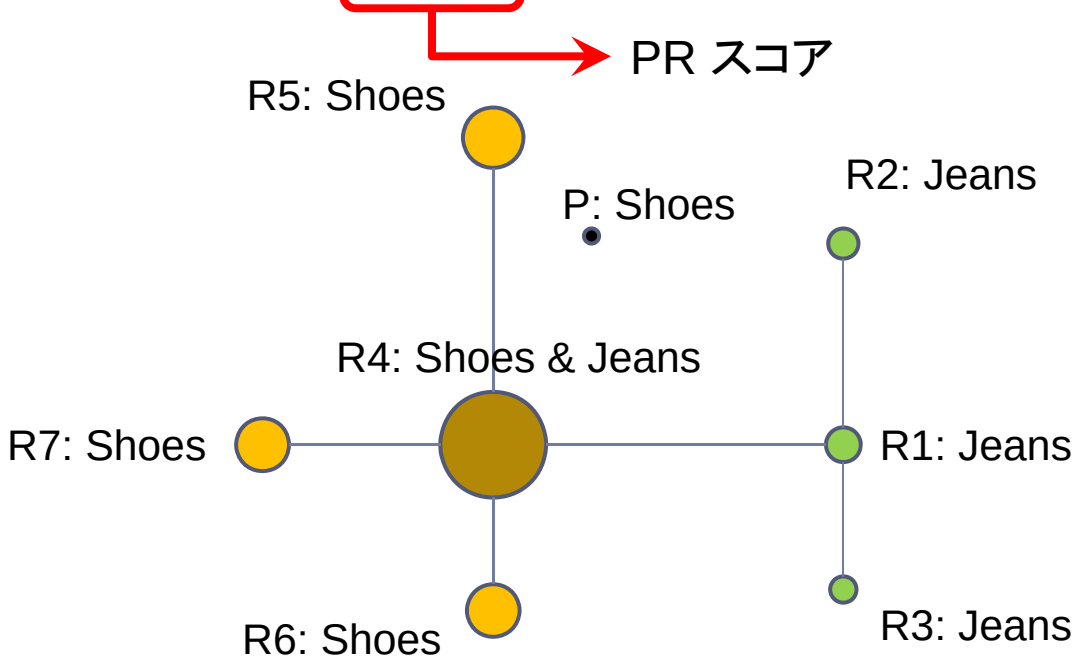
Figure 2: Most influential data objects.

Identifying the Most Influential Data Objects with Reverse Top-k Queries (Vlachou et al.)

- SB 法の問題
  - SB 法は Most Influential Object の個数 $f_I^k()$ が多くなれば，膨大な数の逆 Top-k Query の処理が必要
  - SB 法はすべてのオブジェクトの逆 Top-k Query の処理が終了する必要有
  - インクリメンタルに処理が出来れば高効率 (Branch-and-bound Algorithm (BB))
    - すべての逆 Top-k Query 処理が不要な分だけ高効率

Retrieving Top-k Prestige-Based Relevant Spatial Web Objects (Cao et al.)

- 新しい形の Top-k Queries
  - Location-aware top-k Prestige-based Text Retrieval
    - 例) 位置および評判に基づいた店舗検索
      - 要求に距離的に一番近いのは R5
      - が，R4 のほうがより魅力的（周辺に同種の店舗が多く一番近い）



Retrieving Top-k Prestige-Based Relevant Spatial Web Objects (Cao et al.)

- Location-aware top-k Prestige-based Text Retrieval
  - PR スコア (=personalized PageRank) の計算が非常に高コスト
    - グラフが非常に大きくなる（＝ノード数が大）ため
    - あらかじめ personalized PageRank vector の計算をすることが非効率であるため
    - 空間的距離やすべてのノードの PR スコアを計算すること自体が無駄であるため
  - 既存の Personalized PageRank 計算法で十分では?!

Retrieving Top-k Prestige-Based Relevant Spatial Web Objects (Cao et al.)

- Location-aware top-k Prestige-based Text Retrieval
  - グラフの特徴を考えると Web/entity-relation graph とは相違，つまりグラフの特徴を考えた Personalized PageRank の計算法が必要
    - 類似する店舗は近隣に配置，つまり部分グラフを形成する傾向がある
    - 部分グラフ内のノード数は少，すなわち部分グラフのサイズはそれほど大きくない
    - 方法としては二通り（特殊なグラフに対する Personalized PageRank の計算法）
      - ES-EBC
        - グラフ全体のPRスコアを計算せず，PRスコアを計算するノードの個数に制限をかける
      - S-EBC
        - PRスコアを計算する部分グラフを決定する