VLDB 2015
R22-2, R16-4

担当: 趙　菁(名大)

Some figures are copied from VLDB 2015 proceedings.

担当論文

- 一本目：Searchlight: Enabling Integrated Search and Exploration over Large Multidimensional Data
  - Alexander Kalininら(Brown University)

- 二本目：Reliable Diversity-Based Spatial Crowdsourcing by Moving Workers
  - Peng Chengら (HKUST)

Searchlight: Enabling Integrated Search and Exploration over Large Multidimensional Data

- 研究問題
  - データ探索 ⇨ 検索問題

  何を探すことが分かるが，どこで探すことは問題になる
- 検索問合せ (データセット：SDSS [1])
  - First-order 問合せ：単一な領域
    - Q1: [2,5]° by [3,10] °において，r-等級星の平均値が12以内の領域
  - High-order 問合せ：複数の領域
    - Q2: 異なる部分に存在し，等級の差が1以内の二つの天体領域
  - Optimization 問合せ：最適な領域
    - Q3: 2°by3 °において，r-等級の値が最小な天体を含める領域

[1] SDSS：The Sloan Digital Sky Survey. 天体情報を含めるデータセット

問題点

- 伝統のDBMSは複雑な問合せに支持しない
- 最先端の検索技術は大規模な外部メモリデータセットに適用できない

システム概観

- 統合的な検索と問合せ
  - CP Solver：高度な検索
  - DBMS 機能：データ格納と問合せ処理
- 分散的かつスケーラブルな操作
  - マルチコア機械：効率的な分散と並列操作
- 実装
  - SciDB（多次元配列DBMS）
  - Google’s Or-Tools（CP Solver を含める）

CP Solver

- Constraint Programming (CP) solver
  - 組合せ検索問題に対応

1. 問題をモデリング
   - 変数とドメイン，制約を設定（ユーザ定義）
2. 問題を解決
   - 探索手法（backtracking searchなど）を用いて、解を見つける

統合的な問合せ処理

- 二段問合せ処理
  - Solver ： 投機的な実行 (speculative execution)
    - シノプシス（synopsis） を用いて，
      近似的にAPI callsに答える
  - Validator：結果を検証
    - オリジナルデータを用いる

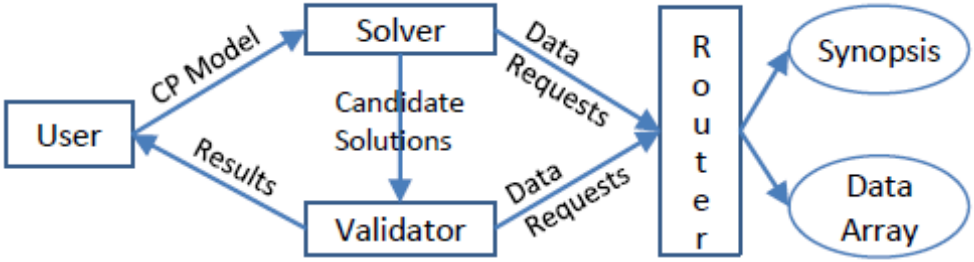


Figure 1: Two-level search query processing.

分散処理

- 問合せ処理の二段階とも分散
  - Solvers: 分散的な検索空間
  - Validators: 対応するノードに割り当てされる
  - 数は任意、同じノードにいる必要なし

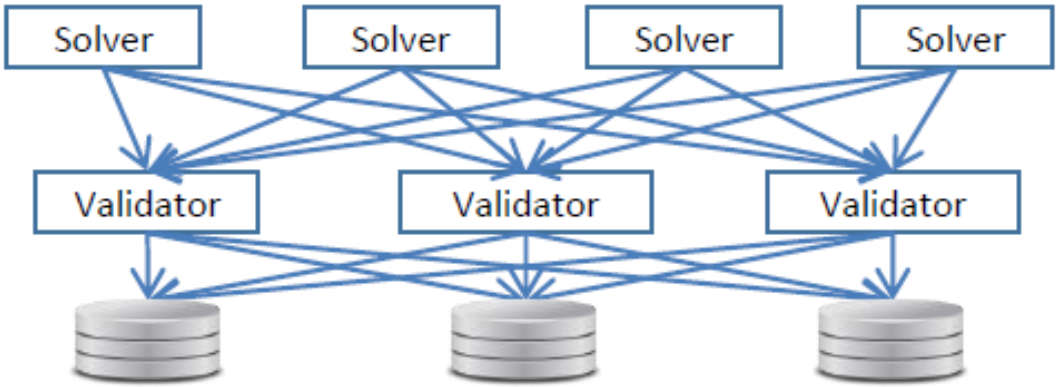


Figure 3: Distributed Searchlight with Solver and Validator layers.

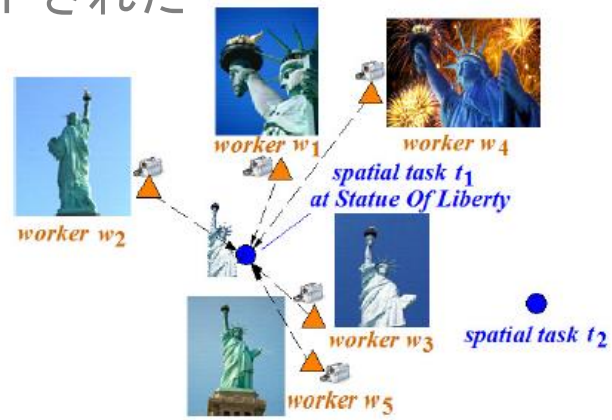
流れ

- 一本目：Searchlight: Enabling Integrated Search and Exploration over Large Multidimensional Data
  - Alexander Kalininら(Brown University)
- 二本目：Reliable Diversity-Based Spatial Crowdsourcing by Moving Workers
  - Peng Chengら (HKUST)

Reliable Diversity-Based Spatial Crowdsourcing by Moving Workers

- 対象：空間クラウドソーシング
  - Moving sensors: モバイルデバイスを用いた人間
  - Sensing tasks: 写真を撮る，ビデオを記録など
- 問題点
  - データの信頼度 (Reliability)
    - 誤った答え
      - 例：偽造の写真/ビデオはアップロードされた

  - 時空間多様性 (Spatial-temporal diversity)
    - 例：ある建物の写真を撮る
      - 異なる角度 → 複数の景色
      - 異なる時刻 →豊かな情報



問題定義｜信頼度

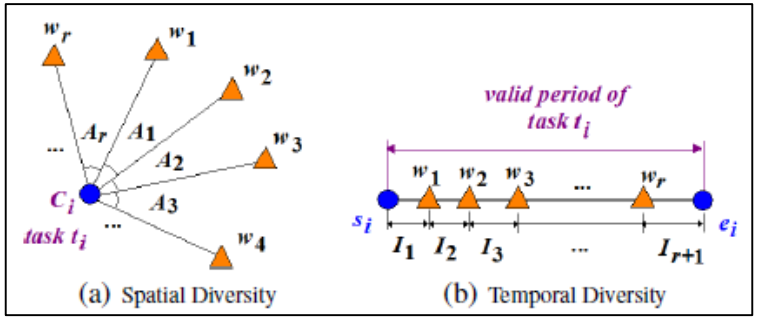
信頼度はタスクを完成できるワーカが
存在するかどうかを示す

- 信頼度: $rel(t_i, W_i) = 1 - \prod_{\forall w_j \in W_i}(1 - p_j)$
  - t_i: タスク
  - W_i: 割り当てされるワーカ集合
  - p_j: ワーカw_jが正確的にタスクを完成する確率

問題定義｜多様性

エントロピーを用いて、時間と空間の多様性を表現

- Spatial Diversity (SD)
  - $SD(t_i) = -\sum_{j=1}^{r} \frac{A_j}{2\pi} \cdot \log(\frac{A_j}{2\pi})$
- Temporal diversity(TD)
  - $TD(t_i) = -\sum_{j=1}^{r+1} \frac{I_j}{e_i - e_s} \cdot \log(\frac{A_j}{e_i - e_s})$
- Spatial-Temporal Diversity (STD)
  - $STD(t_i, W_i) = \beta \cdot SD(t_i) + (1 - \beta) \cdot TD(t_i)$



問題定義 | RDB-SC problem

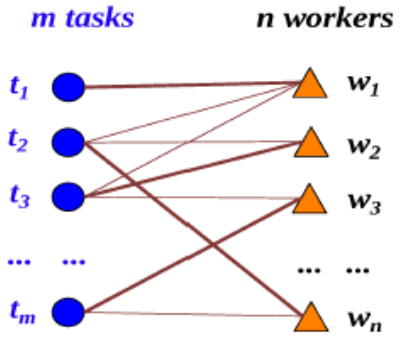
- RDB-SC problem
  (Reliable Diversity-Based Spatial Crowdsourcing):
  各タスク $t_i \in T$ にワーカ集合 W_i を割り当て：
  1. 時間の**有効性**
  2. タスクの信頼度 $min_{i=1}^{m} rel(t_i, W_i)$ が**最大化**され
  3. 時空間多様性の期待値 $E(STD(t_i))$ の
     総和 $total\_STD$ が**最大化**され

提案手法

- 近似手法
  - Greedy Approach
  - Sampling Approach
  - Divide-and-Conquer
- 索引構造
  - RDB-SC-Grid

近似手法

- Greedy approach
  - 非減少プラパティー
  - 繰り返してワーカをタスクに割り当て、かつ高いスコアを求める
- Sampling approach
  - K 個のランダムサンプルを抽出し、スコアが一番高いサンプルを返す
- Divide-and-conquer approach
  1. 繰り返してRDB-SC 問題を二つのサブ問題に分ける
  2. サブ問題を解決する（Greedy or sampling algorithm）
  3. サブ問題の結果を統合する



エッジ：時間の有効性

索引構造(RDB-SC-Grid)

- $1/\eta^2$ 正方形のセル
  - $\eta < 1$, decided based on a cost model(Appendix H)
- セル内容:

ID	Task list	Worker list	Bounds	Tcell_list
$cellid$	(tid, l, s, e)	$(wid, l, v, \alpha^-, \alpha^+, \mathrm{p})$	$[v_{min}, v_{max}]$ $[\alpha_{min}, \alpha_{max}]$ $[s_{min}, e_{max}]$	少なくとも一人のワーカが到着できるセルのIDs

タスクの有効期間

信頼度

移動方向の角度の範囲

タスクの時間制約

- 枝刈り方針
  - **到着不可能なセル**をフィルタする