【VLDB 2015勉強会】

Session 21-2: Spatial Databases

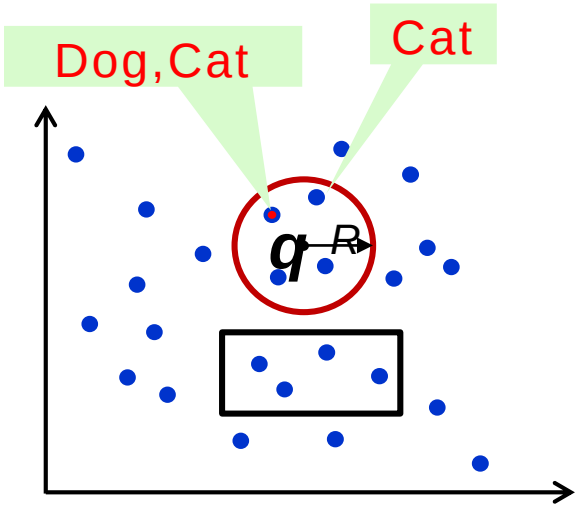
担当：胡（名大）

Some figures are copied from VLDB 2015 proceedings.

Selectivity Estimation on Streaming Spatio Textual Data Using Local Correlations

Xiaoyang Wang(University of New South Wales), Ying Zhang(University of Technology Sydney), Wenjie Zhang, Xuemin Lin, Wei Wang(University of New South Wales)

- 従来のSpatial-keyword search問題との区別
- タイトルのSelectivity Estimationの意味

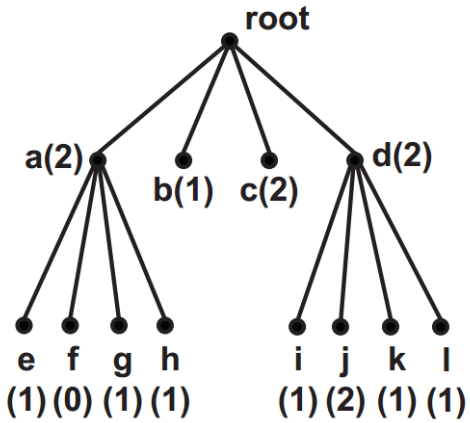


中心***q***, 半径*R*の円(矩形)内で
qのkeywordをいずれも当たる
オブジェクトを検索

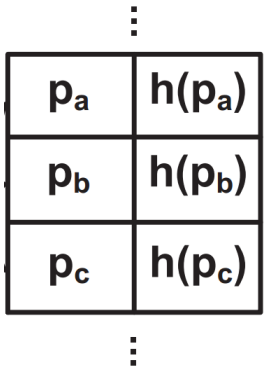
- 従来のSpatial-Keyword問題:　地図上の***q＝{{Dog,Cat},{座標}}***を検索し，すべての結果をユーザに送り出す データベース化は必要
- 本文のSelectivity Estimation問題:　地図上のqを検索し，条件を満たす結果の合計数だけをユーザに送り出す データベース化はしない

Core-technique

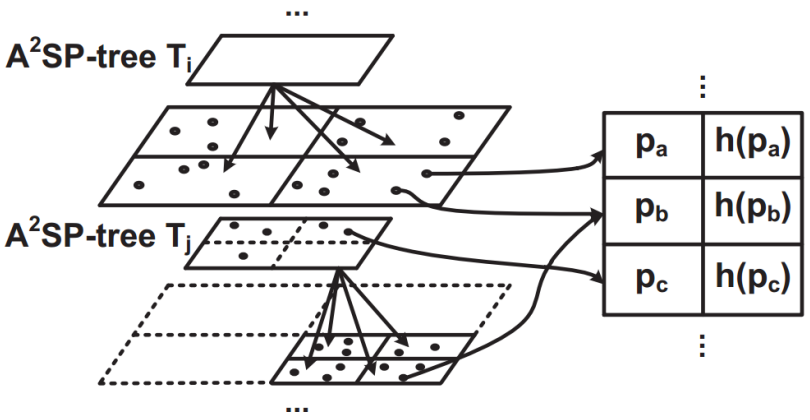
- 本論文で使った三つのテクニック
  - Adaptive Spatial Partition Tree(Quad treeのバリエーション)
  - K minimal values (hash関数を使って集合の大きさを予測する)
  - Bayesian Network (Keyword間の潜在的な関連を発見する)
- いまSpatial-keyword系の研究にはtree系とsignature系の研究の組合せは常態
- 本論文としては, ASPはtree系, KMVはsignature系, Bayesian Networkは新しいアイデア



Spatial側のQuad-tree索引



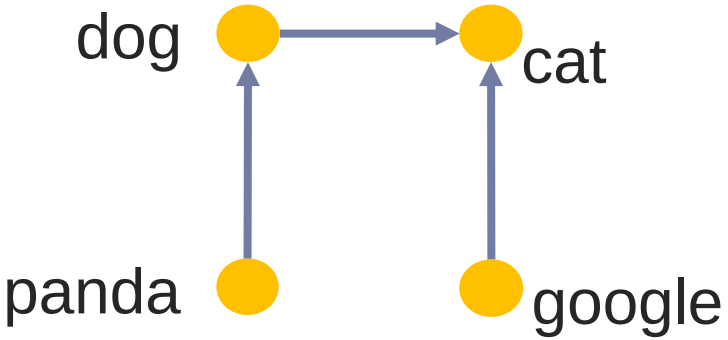
Keyword側の signature hashtable



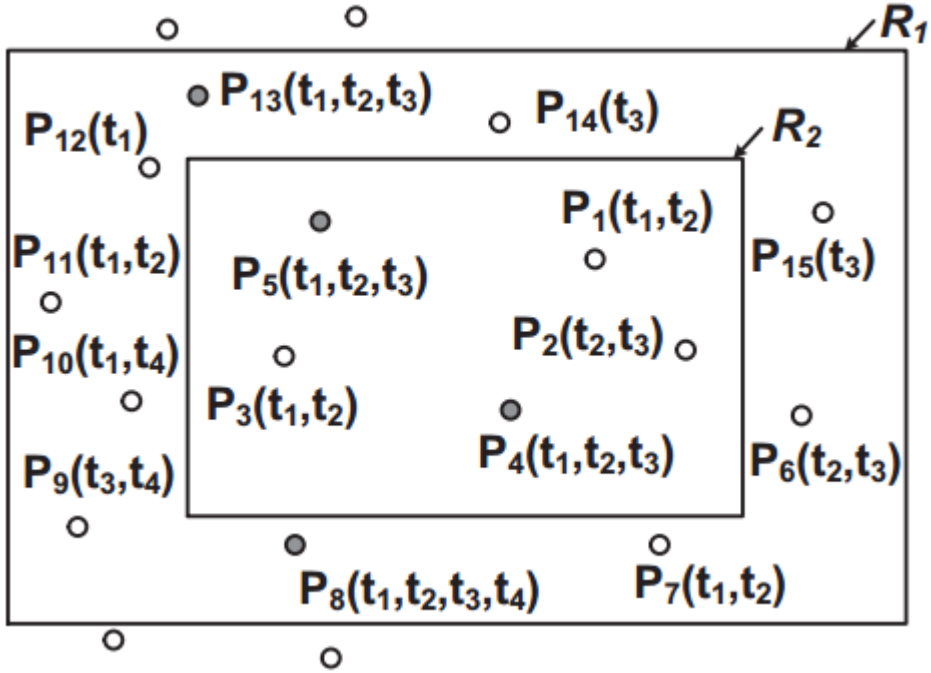
signatureとQuad-treeの組合せ

Bayesian Network

- Machine Learningの話
- t_1=cat
- t_2=dog
- t_3=panda
- t_4=google
- cat,dog,pandaの関連度が高い



機械学習の結果



範囲R1でdog,cat,pandaの文字列が同時に出現したレコードの合計数の予測値の計算は

$$\hat{A} = N \times P_{B1}(T_1 = \text{dog} | T_2 = \text{cat}) \times P_{B1}(T_3 = \text{panda} | T_2 = \text{cat}) \times P_{B1}(T_2 = \text{cat})$$