Algorithms for Mining Distance-Based Outliers in Large Datasets

Knorr, Edwin M.
Ng, Raymond T.
1998

Outliers

„An outlier is an observation that deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism”

- Douglas M. Hawkins

Problem

- Finden von Outliers
- Effizienz
- Dimensionen

Existierende Lösungen

- Sehr viele
- Distribution-based ($k = 1$)
- Depth-based ($k \leq 2$)
- Clustering-Algorithmen

Lösung

- Distance-based outlier detection
- $DB(p, D)$
- 4 Algorithmen

Beispiel

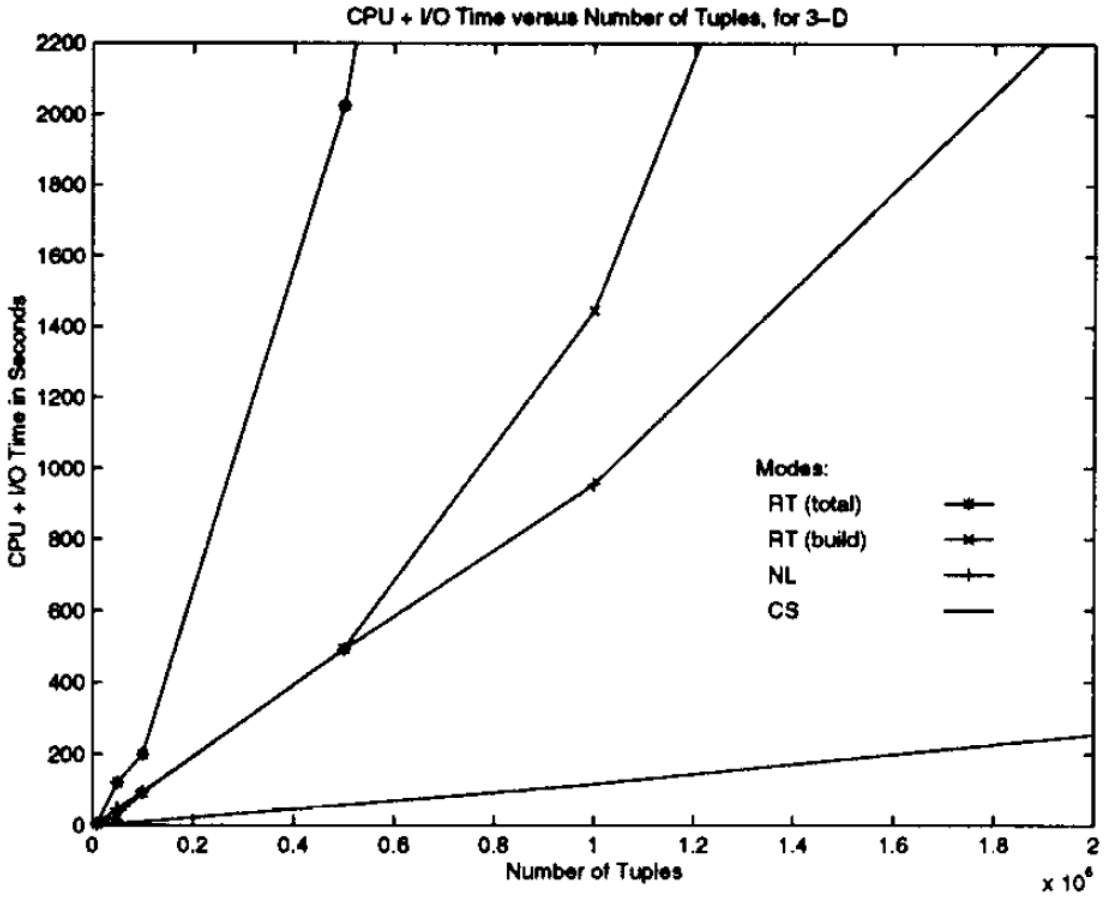
Algorithm FindAllOutsM

1. For $q \leftarrow 1, 2, \ldots m, \quad Count_q \leftarrow 0$
2. For each object P, map P to an appropriate cell C_q, store P, and increment $Count_q$ by 1.
3. For $q \leftarrow 1, 2, \ldots, m, \quad$ if $Count_q > M$, label C_q *red*.
4. For each *red* cell C_r, label each of the L_1 neighbours of C_r ***pink***, provided the neighbour has not already been labelled *red*.
5. For each non-empty *white* (i.e., uncoloured) cell C_w, do:
   a. $Count_{w2} \leftarrow Count_w + \sum_{i \in L_1(C_w)} Count_i$
   b. If $Count_{w2} > M$, label C_w ***pink***.
   c. else
      1. $Count_{w3} \leftarrow Count_{w2} + \sum_{i \in L_2(C_w)} Count_i$
      2. If $Count_{w3} \leq M$, mark all objects in C_w as outliers.
      3. else for each object $P \in C_w$, do:
         i. $Count_P \leftarrow Count_{w2}$
         ii. For each object $Q \in L_2(C_w)$, if $dist(P, Q) \leq D$:
            Increment $Count_P$ by 1. If $Count_P > M$, P cannot be an outlier, so goto 5(c)(3).
         iii. Mark P as an outlier.

Rechtfertigung

- Komplexitätsbetrachtungen
- Messungen
- Variierung von
  - Dimensionen
  - Datensetgröße

Ergebnispräsentation



Ergebnispräsentation

N	CS	NL	KD
20000	0.32	1.02	3.14
40000	0.54	4.26	20.49
60000	0.74	9.64	33.08
80000	1.04	17.58	54.66
100000	1.43	27.67	104.28

Anwendungsgebiete

- E-Commerce
- Kreditkartenbetrug
- Leistungsanalyse von professionellen Athleten

Checkliste

- Stimmt das Resultat?
- Erkenntnisgewinn?
- Neue Ideen?
- Problem wichtig?
- Ergebnis relevant?

Kritikpunkte

- Pseudo-Code
- Datenset
- Ansonsten…