Ludwig Maximilians Universität München
Institut für Informatik
Lehr- und Forschungseinheit für Datenbanksysteme

Skript zur Vorlesung

Datenbanksysteme II

Sommersemester 2005

Kapitel 8: Hochdimensionale Räume

Vorlesung: Christian Böhm
Übungen: Elke Achtert, Peter Kunath

http://www.dbs.informatik.uni-muenchen.de/Lehre/DBSII

Datenbanksysteme II
Kapitel 8: Hochdimensionale Räume

Inhalt

Einführung

- Anfrageleistung von Indexstrukturen verschlechtert sich mit zunehmender Dimension
  "Curse of Dimensionality"
  → Häufig haben scanbasierte Methoden bessere Anfrage-Performanz als z.B. R*-Bäume
- In diesem Kapitel:
  - Ermittlung der Ursachen mit Hilfe eines Kostenmodells
  - Optimierung der Indexstrukturen sowie der Algorithmen zur Anfragebearbeitung
  - Entwicklung neuer Indexstrukturen, die besonders an die Problemstellung hochdimensionaler Datenräume angepasst sind

Inhalt

Kostenmodell für R-Bäume

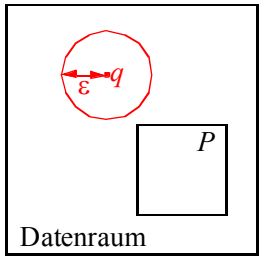
([Berchtold S., Böhm C., Keim D., Kriegel H.-P.: *A Cost Model for Nearest Neighbor Search in High-Dimensional Data Spaces,* PODS 1997])

Ziel: Schätzung der zu erwartenden Anzahl der Seitenzugriffe bei der Anfragebearbeitung

- für R-Bäume und verwandte Indexstrukturen
- verschiedene Anfragetypen:
  - Bereichsanfragen
  - nächste-Nachbar-Anfragen sowie *k*-nächste-Nachbar-Anfragen
  - verschiedene Metriken, hier nur euklidische und Maximums-Metrik (Ellipsoid-Anfragen sind schwierig zu modellieren)
- Einschränkungen:
  - idealisierte Indexstruktur:
    überlappungsfrei
    Seitenregionen sind annähernd quadratisch (“so quadratisch wie möglich”)
  - Datenraum ist der Einheits-Hypercube $[0..1]^d$
  - *zunächst: Punkte und Anfragen folgen einer unabhängigen Gleichverteilung*
  - *später: Beschreibung der Datenverteilung durch fraktale Dimension*
    (genauere Darstellungsmethoden der Datenverteilung wie z.B. Histogramme sind im Hochdimensionalen schwierig)

Kostenmodell Bereichsanfragen (1)

- Bekannt:
  - Radius ε der Anfrage
  - Ausdehung der Seitenregion
    → später wird beides geschätzt werden
- Unbekannt:
  - relative Lage von Seitenregion und Zentrum der Anfrage (*Anfragepunkt*)
  - beides wird als unabhängig gleichverteilt angenommen, d.h. jede Position von Anfragepunkt und Seitenregion ist gleich wahrscheinlich
- Gesucht:
  - Wahrscheinlichkeit, mit der die Query *q* auf die Seite *P* zugreift (Zugriffswahrscheinlichkeit)
  - entspricht Wahrscheinlichkeit, mit der sich der Kreis mit der Seitenregion schneidet



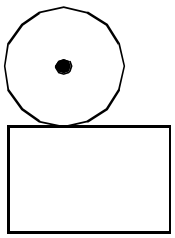
Kostenmodell Bereichsanfragen (2)

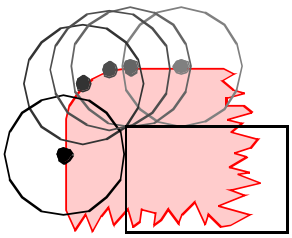
- Das Problem ist leicht zu lösen, wenn z.B. die Anfrage punktförmig ist:

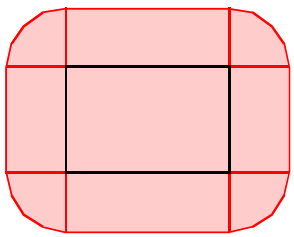
$$\text{Zugriffswahrscheinlichkeit} = \frac{\text{Volumen Seitenregion}}{\text{Volumen Datenraum}}$$

- ebenso bei punktförmiger Seitenregion
  → Wahrscheinlichkeitsrechnung (Kombinatorik) benötigt *punktförmige* Ereignisse
- Trick um punktförmige Ereignisse zu erhalten:
  Transformiere Bereichsanfrage in eine äquivalente Punktanfrage:
  - Verkleinere die Bereichsanfrage zum Punkt
  - Vergrößere in gleichem Maß die Seitenregion
  - so dass die neue Punktanfrage auf die vergrößerte Seitenregion zugreift gdw. die Bereichsanfrage auf die ursprüngliche Seitenregion zugreift

Kostenmodell Bereichsanfragen (3)



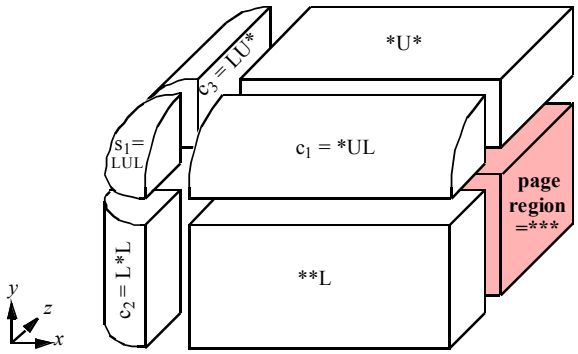




- Das entstehende Objekt heißt *Minkowski-Summe* von Anfrage- und Seitenregion:
  - ursprüngliches Rechteck
  - an jeder Kante ist ein Quader der Breite ε angehängt
  - an jeder Ecke ist ein Viertelkreis mit Radius ε angehängt

Kostenmodell Bereichsanfragen (4)

- Im dreidimensionalen Fall (Grafik unvollständig):



  - ursprünglicher Quader: 3-dimensional rechteckig, 0-dimensional rund
  - an jeder Oberfläche: Quader mit Grundfläche wie Oberfläche, Dicke ε
  - an jeder Kante: ¼ Zylinder: Länge wie Kante, Grundfläche ist Kreis mit Radius ε
  - an jeder Ecke: ⅛ Kugel mit Radius ε

Kostenmodell Bereichsanfragen (5)

- Ein *d*-dimensionaler *Hypercube* hat neben Ecken (0-dimensional), Kanten (1d) und Flächen (2d) auch noch 3-dimensionale, 4-dimensionale .... (*d* - 1)-dimensionale "Oberflächen"-Segmente (engl. *faces*)
  - an jedem *i*-dimensionalen Segment hängt ein Objekt ("*Hyperzylinder*"), das in *i* Dimensionen würfelförmig ist und in (*d* - *i*) Dimensionen kugelförmig ist (genau genommen der 2^{d-i}-te Teil einer solchen Hyperkugel)
- Wie viele Ecken, Flächen, ... *i*-dimensionale Segmente hat ein *d*-dimensionaler Würfel?
  Hierzu führen wir eine Notation ein, die, jedem Oberflächensegment (incl. dem urspr. Hypercube) ein *d*-Tupel über dem Alphabet aus den drei Symbolen L, U, und * zuordnet.
  Hierbei bedeutet:
  - L die untere Grenze (lower bound) in einer Dimension
  - U die obere Grenze (upper bound) in einer Dimension
  - * den gesamten Bereich zwischen der unteren Grenze und der oberen Grenze

Kostenmodell Bereichsanfragen (6)

- Beispiel:
  - Enthält ein Tupel kein *, so bezeichnet es eine *Ecke* (z.B. LUL die linke, obere, vordere, Ecke des dreidimensionalen Würfels)
  - Enthält ein Tupel genau ein *, so bezeichnet es eine *Kante* (z.B. LU* die Kante links oben, von vorne nach hinten)
  - Ein Tupel mit *i* Sternchen bezeichnet ein *i*-dimensionale Oberflächensegment
  - Das Tupel, das *nur* aus *d* Sternchen besteht, bezeichnet den originalen Hypercube
- Anzahl der Tupel mit *i* Sternchen:

$$\binom{d}{i} \cdot 2^{d-i}$$

$\binom{d}{i}$: verteile *i* Sternchen über *d* Positionen

2^{d-i}: fülle die verbleibenden *d-i* Positionen mit L oder U

Kostenmodell Bereichsanfragen (7)

- Gesamte Formel für das Volumen der Minkowski-Summe aus Hypercube mit Seitenlänge *a* und Kugel mit Radius ε:

$$V_{Mink}(a,\varepsilon) = \sum_{0 \le i \le d} \binom{d}{i} \cdot 2^{d-i} \cdot a^i \cdot \frac{V_{(d-i)-\dim.Kugel}(\varepsilon)}{2^{d-i}} = \sum_{0 \le i \le d} a^i \cdot V_{(d-i)-\dim.Kugel}(\varepsilon)$$

- Das Volumen einer *j*-dimensionalen Kugel lässt sich folgendermaßen ermitteln:

$$V_{j-\dim.Kugel} = \frac{\pi^{j/2} \cdot r^j}{\Gamma(j/2+1)}$$

wobei Γ die Gamma-Funktion (Erweiterung der Fakultät in reelle Zahlen) darstellt, mit:

$$\Gamma(x+1) = x \cdot \Gamma(x) \qquad \Gamma(1) = 1 \qquad \Gamma(1/2) = \sqrt{\pi}$$

- Mit V_{Mink} kann die Zugriffswahrscheinlichkeit einer einzelnen bekannten Seite bereits ermittelt werden. Dies werden wir später bei einer Optimierungstechnik anwenden.
- Im allgemeinen interessieren die Kosten für den gesamten Index; Summation der Zugriffswahrscheinlichkeiten aller Seitenregionen zu teuer

Kostenmodell Bereichsanfragen (7)

Schätzung der Seitenlänge des Hypercube

Benutze eine durchschnittliche Seite anstatt der konkreten Seiten.

Für jede Indexebene i läßt sich die Anzahl der Seiten n_i ermitteln, sofern die durchschnittliche Speicherauslastung (su_{eff}) bekannt ist (aus Data Dictionary):

Sei $C_{\text{eff}} := C \cdot su_{\text{eff}}$ *die effektive Kapazität der Seiten (durchschnittliche Anzahl in einer Seite gespeicherter Einträge),* N die Gesamtzahl von Featurevektoren

- $n_0 := N/C_{\text{eff,data}}$ (Anzahl der Datenseiten)
- $n_i := n_{i\text{-}1}/C_{\text{eff,dir}}$ (Anzahl der Seiten auf Directory-Ebene i)

Annahmen:

- Seitenregionen haben Volumen $1/n_i$ (1 ist das Volumen des Datenraums $[0..1]^d$)
- Seitenregionen sind annähernd (hyper-) würfelförmig.

Schätzwert für die Kantenlänge a_i einer Seitenregion auf Indexebene i: $a_i = \sqrt[d]{1/n_i}$

Kosten durch Addition der Zugriffswahrscheinlichkeiten aller Seiten auf allen Ebenen: $\#\text{Zugriffe}(\varepsilon) = \sum_i n_i \cdot V_{Mink}(\sqrt[d]{1/n_i}, \varepsilon)$