Evropský sociální fond
Praha & EU: Investujeme do vaší budoucnosti

Katedra softwarového inženýrství, Fakulta informačních technologií,
České vysoké učení technické v Praze

VYHLEDÁVÁNÍ NA WEBU A V MULTIMEDIÁLNÍCH DB (BI-VWM)

Projekt V - 1

INDEXOVÁNÍ – R-STROM

RNDr. David Hoksza, PhD.

ZADÁNÍ

Cílem projektu je vytvoření vlastní perzistentní implementace R-stromu.

VSTUP

Dotaz nad databází 2D nebo 3D objektů.

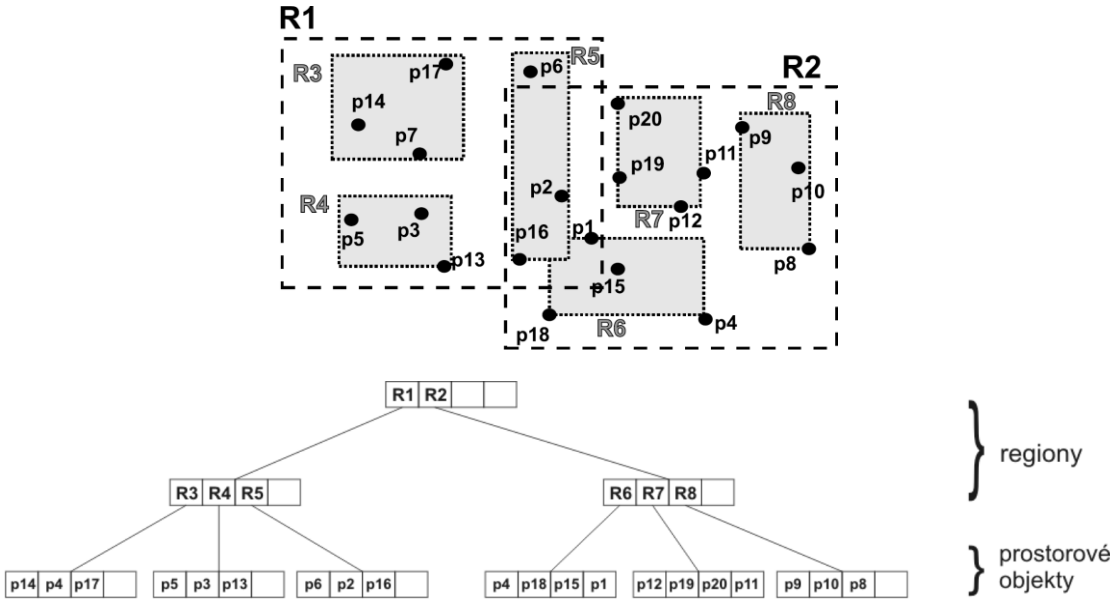
VÝSTUP

Seznam objektů odpovídající dotazu.

INFORMACE/POTŘEBNÉ ZNALOSTI

R-strom je datová stromová struktura využitelná pro vyhledávání n-dimenzionálních objektů. R-strom je často využíván v geografických informačních systémech a je tedy dobře aplikovatelný např. pro vyhledávání na webu v mapových systémech.

R-strom má stromovou strukturu, kde každý uzel má proměnlivý počet záznamů. Záznamy tvoří tzv. MBR (minimum bounding rectangle) regiony, což jsou (potenciálně se pronikající) n-dimenzionální (v mapových systémech n=2) obdélníky hierarchicky dělící prostor.



Vyhledávání objektu v R-stromu probíhá stejně jako v B-stromu až na skutečnost, že při vyhledávání jednoho objektu, může být prohledáno více podstromů. Tato skutečnost je způsobena faktem, že regiony se můžou navzájem překrývat. Leží-li tudíž objekt v překryvu, pak nelze rozhodnout, v kterém z podstromů se objekt skutečně nachází a musí být prohlédnuty všechny podstromy, které se překryvu účastní (příkladem toho může být vyhledání bodu p2 v obrázku - musíme se podívat do R1 i R2). Vyhledáváme-li pomocí rozsahového dotazu, tj. nevyhledáváme jeden konkrétní bod, ale všechny body ležící ve vymezeném prostoru (rozsahu), pak při procházení stromem musíme projít vždy všechny podstromy reprezentované pomocí MBR, které mají neprázdný průnik s daným rozsahem.

Pro rozhodnutí, kam vložit objekt je důležitá minimalizace následujících veličin:

- *Pokrytí* úrovně stromu, tj. celková plocha regionů v dané úrovni stromu.

- *Přesah* na úrovni stromu, tj. celková velikost průniků ploch regionů.

Minimalizujeme-li pokrytí, pak tím také minimalizujeme tzv. *mrtvou plochu*, tj. plochu, která neobsahuje objekty. Tím se snižuje celkový prohledávaný prostor, který jistě neobsahuje objekty. Minimalizujeme-li na druhé straně přesah, pak snižujeme pravděpodobnost prohledávání více podstromů, jelikož je menší pravděpodobnost, že hledaný objekt bude zasahovat do překrývajících se regionů. Na základě znalosti faktů o pokrytí a přesahu můžeme zajistit vkládání takové, které bude tyto veličiny minimalizovat. Tedy listový uzel, kam vložit objekt, je nalezen tak, že je procházen strom od kořene a v každém uzlu se při rozhodnutí, do kterého potomka bude objekt vložen, řídíme pravidlem, že je vybrán takový uzel, který potřebuje nejmenší rozšíření, pokud do něj bude nový objekt vložen. V případě, že tomuto kritériu vyhovuje více uzlů, je vybrán ten, jehož výsledná plocha bude nejmenší. Takto rekurzivně dojdeme až do listu. Pokud není zaplněn, pak je do něj objekt vložen. V opačném případě dochází na dělení regionu. Dělit region lze mnoha způsoby, a jelikož dělení prostoru je pro efektivitu vyhledávání důležité, je třeba vybrat dělení co možná nejlepší, ovšem s ohledem na časovou složitost této operace. Uveďme 3 nejznámější způsoby dělení uvedené Guttmanem:

- *Úplný algoritmus* procházející všechny možnosti
- *Kvadratický algoritmus* - rozděluje objekty postupně do dvou od sebe maximálně vzdálených skupin (nových regionů)
- *Lineární algoritmus* - princip je stejný jako u kvadratického algoritmu, pouze se mění způsob vybrání objektu, který bude v daném kroku přiřazen k jedné z vytvářených skupin

STAVBA APLIKACE

Aplikace by měla obsahovat:

- Persistentní implementace R-stromu.
- Možnost vkládání nových objektů.
- Dotazovací modul s implementací dotazu na nejbližšího souseda a rozsahového dotazu (tj. které objekty náleží do dotazového regionu).

POZNÁMKY K ŘEŠENÍ

Projekt je koncipován jako implementace algoritmu, proto rozhraní může být i textové (nicméně mělo by být relativně ověřitelné korektní fungování algoritmu).

Algoritmus by měl být perzistentní. Tedy by měl být schopen pracovat i s databázemi, které se nevejdou celé do paměti.

DATA

Data lze vygenerovat náhodná.

EXPERIMENTY

V tomto projektu se nabízejí experimenty na testování zrychlení při použití indexu oproti sekvenčnímu průchodu. Dále lze testovat vliv různých parametrů na rychlost, např. maximální velikost uzlů apod.

ZDROJE

- Přednáška *Podobnostní dotazy, agregační operátory.*
- Antonin Guttman. *R-Trees: A Dynamic Index Structure for Spatial Searching,* Proc. 1984 ACM SIGMOD International Conference on Management of Data, pp. 47-57