OLAP

OLAP: egy tisztított és előfeldolgozott adatokra épülő döntéstámogató rendszer, amelynek fő célja a többdimenziós adatokra vonatkozó statisztikai jellegű kérdések gyors megválaszolása és a válaszok szemléletes megjelenítése.
Codd-szabályok az OLAP definiálásához:

1. Multi-dimenzionális nézet
2. Transzparencia (itt most technikai részletek ismerete nélküli könnyű elérhetőség, tehát :áttekinthetőség értelemben)
3. Elérhetőségek (jogosultságok) beállíthatósága
4. Állandó riportozási (lekérdezési) teljesítmény
5. Kliens-szerver architektúra
6. Általános dimenzió fogalom
7. Dinamikus ritka-mátrix kezelés (ez a multidimenzionális modell tárolására vonatkozik, megvalósításra megkötés)
8. Több konkurens felhasználó támogatása
9. Korlátozás nélküli dimenzióműveletek
10. Intuitív adatkezelés (a végfelhasználó számára)
11. Rugalmas riportozás (vagyis beszámoló-készítés, lekérdezés)
12. Korlátlan dimenziószám és aggregációs szint szám

FASMI:

- Gyors(Fast): A rendszer interaktív működésű, vagyis rövid időn belül választ kell adnia.
- Elemzési képesség(Analysis): Az OLAP rendszer lehetőséget teremt változatos, dinamikusan összeállított üzleti és statisztikai elemzések elvégzésére.
- Többfelhasználós környezet(Shared): Az adatokat több felhasználó párhuzamosan használhatja (biztonság)
- Multidimenzionális(Multidimensional): többdimenziós adatábrázolás és adatkezelés
- Információ(Information): az OLAP rendszer tartalmazza az összes közvetlen és származtatott adatot a komplex számítások elvégzéséhez.

Műszaki problémák az OLAP rendszerekben

A konzisztens, naprakész, történeti forrásinformációk beszerzése, multidimenzionális tárolási és ábrázolási módszerek, megfelelő teljesítmény biztosítása

A multidimenzionális adatmodell prezentálása és tárolása

Az adatkocka a multidimenzionális adatok tárolási modelljétől független adatelemzési egysége, ROLAP-, MOLAP-, HOLAP-tárolás

OLAP kocka: a kocka oldalai mentén tüntetjuük fel a dimenziókat, míg a dimenziók által meghatározott metszéspontokban található cellák a tényadatokat tartalmazzák.

- ROLAP(Relational OLAP): a multidimenzionális adatmodellt relációs adatbázisra képezik le csillag és hópehelysémák használatával → relációs jó, mert a fejlesztők már ismerik, sok eszköz van rá, skálázható…
- MOLAP(Multidimensional OLAP): az adatok tárolására multidimenzionális adatmodell. Ez a modell az egyes elemeket többdimenziós vektorban tárolja, ahol azok közvetlen indexeléssel hozzáférhetők.
- HOLAP(Hybrid OLAP): ötvözik a relációs megközelítés skálázhatóságát a multidimenzionális tárolás gyorsabb adatelérési és számítási képességeivel.

ROLAP vs MOLAP tárhelykezelés: ritka adatok kezelésénlé MOLAP ugyanannyi multidimenziós adat tárolásához akár 2-10-szer kevesebb helyet igényel, mint a ROLAP.

Műveletek az OLAP-kockával

- szeletelés(slice): a hiperkocka egyik dimenzióját rögzítjük, és e rögzített dimenzió értéke mentén hajtunk végre elemzést.
- részkocka-kiválasztás(dice): az eredeti hiperkocka dimenzióit csak a kiválasztott értékek mentén vizsgáljuk. (Tehát a kocka egy kisebb részét vizsgáljuk csak.)

- Lefúrás(drill-down): megnöveli az egyik dimenzió felbontását (éves→negyedéves)
- felgörgetés(roll-up): egy adott dimenzió mentén egy magasabb hierarchiaszintre emelkedünk. (hónap→negyedév)
- elforgatás(pivot): a koordinátatengelyek sorrendjét vagy azok irányát felcseréljük megjelenítési okok miatt. (A vizsgált adatok térbeli elforgatása szemléletesebb eredménymegjelenítés érdekében.)
- további műveletek: átfúrás(drill-across) OLAP kockák közötti elemzés; keresztülfúrás (drill-through): a kocka felépítéséhez használt operatív adatokhoz tudunk hozzáférni.

Az előkalkuláció szerepe

A válaszidő javítása a válaszok előre történő kiszámításával és eltárolásával. A teljes előkalkuláció elvégzése óriási számítási és tárolási kapacitást igényel. Részleges előkalkuláció: nehezen kiszámítható, és gyakran használt aggregátumok előzetes létrehozását kívánja meg.
Elvileg tök jó lenne, ha mindent ki tudnánk előre számítani, és ezt egyszerűen elmenteni, majd visszaolvasni, de annyi variáció van számításokra (OLAP alapelv: interkatívan, bármyel dimenziók összekapcsolásával stb.. képzünk kérdéseket), hogy ez lehetetlen. Azonban nem kell minden kérdésre kiszámítani a válaszokat, csak azokat, amelyek:

- sok cellát érintenek
- gyakran vesznek részt a válaszok előállításában
- bonyolult számításokat tartalmaznak

Az adatbázis-robbanás

Az OLAP-adatbázisok aránytalan mértékben nőnek az előkalkuláció mértékének és a dimenziójuk sázmának növelésével. Tévhitek az adatrobbanással kapcsolatban. Előkalkulált adatok mennyisége ritka tényadatterek felett. Teljes és dimenziónkénti növekedési tényező
Adatbázis-robbanás: viszonylag kis mennyiségű alapadat tárigénye a sokdimenziós feldolgozás során több tízszeresére vagy több százszorosára növekszik.
Growth-factor(GF): (x+x')/x ; x alapadatok számossága, x'az adatbázishoz hozzáadott származtatott adatok számossága. Compound growth factor(CGF): dimenziónkénti növekedési tényező, GF k-adik gyöke.

Az előkalkuláció mértékének helyes megválasztása

Az adatrobbanás elleni védekezés lehetőségei: multidimenzionális tárolás és tömörítés, kisebb dimenziószámú hiperkockák párhuzamos használata, az előkalkuláció mértékének helyes megválasztása

Adatbányászat

Az adatbányászat nagy adathalmazokon végzett önálló tudásfeltárás. Felhasználási területek: kereskedelem, pénzügy, távközlés, orvostudomány
adatbányászat: egy interdiszciplináris tudományterület, amely a nagy adathalmazokból történő tudásfeltárással foglalkozik, vagyis az adatokban rejlő újszerű, nem triviálisan előállítható összefüggések és minták automatikus kinyerését célozza.

Módszertan és architektúra

Az adatbányászati folyamat: a cél kijelölése, az adatok előkészítése, a célnak megfelelő algoritmus kiválasztása, és lefuttatása, az eredmények értékelése és prezentálása. Megalapozottsági, megbízhatósági és érdekességi mértékek hozzárendelése az algorizmusokhoz.
Adatbányászati folyamat:

1. projekt pontos céljának kijelölése: különböző doménekhez különböző algoritmusok
    a. adatbányászati alapfeladatok: társítási szabályok keresése/gyakori elemhalmazok keresése/, eltéréselemzés/többi típus elemeire épít, de itt a lényeg az eltéréseken van/, osztályozás(modellezés)/a mintákat előre ismert kategóriákba sorolása/, csoportosítás(klaszterezés)/a mintákat előre nem ismert kategóriákba sorolása/, komplex és strukturálatlan adattípusok bányászata/web- multimédia és téradatbázisok bányászata; érdekes minták és az összefüggések keresése az adott adatforrás felhasználásával/

2. adatok előkészítése (adatkiválasztás, kinyerés, integrálás, tisztítás, transzformáció)
3. adatfeldolgozás (konkrét, kitűzött célnak megfelelő algoritmust futtatunk az előkészített adatokon)
4. kinyert minták és szabályok értelmezése (gépi úton kinyert minták emberi felhasználásra alkalmatlanok)

- Megalapozottság: milyen gyakran fordulnak elő az adatlhalmazban az adott minta vagy szabály felépítésében részt vevő elemek.
- Megbízhatóság: adott szabály erőssége, jósága. “aki ruhát vesz, az fiatal” nem elég erős
- Érdekesség: triviális információk nem kellenek. “Az emberek 99%-nak 2 mellbimbója van”

Osztályozás

Modellezési feladat, besorolási kategória meghatározása tanítópéldák alapján

Minőségi paraméterek

Módszer jósága(pontossága): osztályozási hiba becslésével mérhető. Túltanulás: ha az osztályozó algoritmus a tanítóhalmaz speciális összefüggéseit és mintáit tanulja meg ahelyett, hogy ezek általánosításával építené fel a modellt. Osztályozás megalapozottsága: az egyes osztályokhoz tartozó tanítóminták száma jellemzi
Teljesítményparaméterek: betanulási idő, futási idő, skálázhatóság. Robosztusság. Eredmények értelmezhetősége.

Osztályozási módszerek

- döntési fák: bonyolult osztályozási döntéseket egyszerűbb döntések sorozataként kezelik
- neurális hálózatok: lineáris szeparációra képes univerzális függvényapproximátorok, amelyek statisztikai modellezőeszközként is felhasználhatók.
- bayes-i osztályozóhálók: Két alapelve
    - Bayes-tétel: ismeretlen feltételes valószínűségek meghatározása imertekkel
    - maximum likelihood elv: alapján a legnagyobb valószínűségű osztály mellett döntünk
- legközelebbi szomszédok módszere: hasonló adatok egymáshoz közel helyezkednek el az attribútumtérben.
- lineáris regresszió: valós értékkészletű változókra alkalmazható statisztikai módszer

Klaszterezés

A klaszterezés olyan többváltozós, struktúrafeltáró elemzés, amelynek segítségével az objektumokat tulajdonságaik hasonlósága alapján felépített csoportokra osztjuk

Minőségi paraméterek

Hasonlóságfüggvény, amelynek alapján az elemek hasonlóságát értelmezhetjük. A klaszterezés feladata nem egyértelmű feladat. Klaszterezés jósága: skálainvariancia, kifejezőkészség, konzisztencia. Teljesítmény, robosztusság, értelmezhetőség.

Klaszterezési módszerek

- particionáló algoritmusok: induláskor az összes elemet besorolják valamilyen halmazokba, majd a pillanatnyi eerdmények fokozatos alakításával, iteratív úton jutnak el a végeredményt képező klaszterekig
    - k-közép algoritmus: kiválasztja a kezdeti klaszterközéppontokat → minden elemet besorol a hozzá legközelebb eső klaszterközponttal jellemzett csoportba → minden lépésben képezzük a klaszter elemeinek vektori átlagát, és ez lesz a klaszter új középpontja
- hierarchikus algoritmusok: az elemeket egymástól való távolságuk alapján egy hierarchikus faszerkezetbe rendezi, ennek alapján alakítják ki a klasztereket
    - felülről lefel(top-down): eleionte minden elem egyetlen klaszterbe kerül, majd 2 kisebb, .. klaszterenként 1 elem.
    - lentről felfele(bottom-up): eleinte minde elem külön klaszterbe, majd egyesít stb.. 1 klaszterben minden elem a végén
- sűrűségalapú algoritmusok: az összetartozó csoportokat az elemek sűrűsége, ill az elemsűrűség változása alapján próbálják megragadni. tipikus megvalósítás: DBSCAN

- rácsalapú algoritmusok: futási sebesség javítása érdekében a sokdimenziós teret egy ráccsal kisebb részekre, cellákra osztják fel → a tényleges elemeket rácspontokra képezi, majd ezzel a reprezentációval dolgozik
- modellalapú algoritmusok: az elemekből levont következtetésekre építve minden egyes klaszterre egy modellt állítanak fel, majd az ezekhez való illeszkedést vizsgálva leresik meg a klaszterekbe leginkább illő elemeket
- spektrális algoritmusok: az elemek hasonlósága (távolsága) alapján képzett hasonlóságmátrix sajátértékeit és sajátvektorait dolgozzák fel annak érdekében, hogy a feladatot egy kisebb dimenziószámú problémává alakítsák

Társítási feladatok

Asszociációs szabyálok kialakítása bevásárlói kosarak tartalmának elemzése alapján. Gyakori elemhalmazok és minták felderítése. Hasznosítás kereskedelmi, kémiai és pénzügyi területeken.

Társítási szabályok leírása és értékelése

MAtematikai formalizmus gyakori elemhalmazokhoz és társítási szabályokhoz. Megalapozottság és megbízhatóság. Hasznosság

Módszerek társítási szabályok keresésére

A keresési tér túl nagy a nyers erő alkalmazásához. A priori elv. Szintenként haladó és mélységi bejárást végző algoritmusok. Vízszintesen és a függőlegesen vizsgálatot végző módszerek

Priori elv: ha egy elemhalmaz gyakori, akkor minden részhalmazának gyakorinak kell lennie.
A priori elv alapján az algoritmusok legyártanak néhány alapvető elemhalmazt, az ún. jelölteket, és először ezek gyakoriságát vizsgálják.
A gyakori elemhalmazokat előállító módszereket 2 csoportra oszthatjuk: mélységi és szélességi bejárást alkalmazó algoritmusok.
A jelöltek gyakoriságának számlálási módját tekintve 2 alapvető módszer: vízszintesen és függőlegesen vizsgáló. A vízszintesen dolgozó algoritmusok a tranzakciókat úgy tárolják, hogy minden tranzakcióhoz feljegyzik, hogy mely termékek építik fel a tranzakciót, a gyakori jelöltek számlálása pedig úgy történik, hogy minden tranzakción sorban végighaladunk, és megvizsgáljuk, melyik jeleöltek szerepelnek benne, majd növeljük a hozzájuk tartozó számlálót. A függőlegesen vizsgáló algoritmusok nem tranzakciónként haladnak, hanem minden egyes gyakori elemhez eltárolják, hogy azok melyik tranzakciókban szerepelnek.