Tietokantajärjestelmät spatiaalisen tiedon louhinnassa

Mikko Valjento

Helsinki 18. huhtikuuta 2003
Esitelmä, Spatiaalisen tiedon louhinnan seminaari
HELSINGIN YLIOPISTO
Tietojenkäsittelytieteen laitos

Tietokantajärjestelmät spatiaalisen tiedon louhinnassa

Mikko Valjento (valjento@helsinki.fi)

Esitelmä
Tietojenkäsittelytieteen laitos
Helsingin yliopisto
18. huhtikuuta 2003, 14 sivua

Esitelmässä käsitellään spatiaalisen tiedon louhinnassa käytettävien tietojärjestelmien rakennetta. Aluksi esitellään yleisille monikomponenttimalliin perustuva arkkitehtuuriratkaisu tiedonlouhintajärjestelmille. Tämän jälkeen käsitellään spatiaalisen tiedon mallinnusta ja käsittelyä spatiaalisissa tietokantajärjestelmissä. Tämän jälkeen käsitellään spatiaalisen tiedon käsittelyn tehostamista spatiaalisella indeksoinnilla. Spatiaalisen tiedon indeksointiin käytetyistä tietorakenteista esitellään R-puu, sekä sen muunnokset R*-puu ja R+-puu. Lisäksi käsitellään spatiaalisessa tiedon louhinnassa käytettyjä naapurustokaavioita (neighbourhood graphs) ja naapurustoindeksejä (neighbourhood indeces). Lopuksi esitellään geo-spatiaalisen tiedon louhintaan kehitetty GeoMiner-tiedonlouhintajärjestelmä.

Aiheluokat (Computing Reviews 1998): E.1 Data Structures, H.2.3 Languages, H.2.8 Database Applications, H.3.3 Information Search and Retrieval

Avainsanat: tietämyksen muodostaminen, tiedon louhinta, spatiaaliset tietokannat

Sisältö

1 Johdanto

Spatiaalinen tieto eli paikkatieto (spatial data) on tietoa, joka on riippuvaista kohteen sijainnista tai muodosta kaksi- tai useampiulotteisessa avaruudessa. Spatiaalisen tiedon louhinnalla (spatial data mining, knowledge discovery in spatial databases) tarkoitetaan kiinnostavan ja yleensä aikaisemmin tuntemattoman tietämyksen löytämistä spatiaalista tietoa sisältävistä tietokannoista.

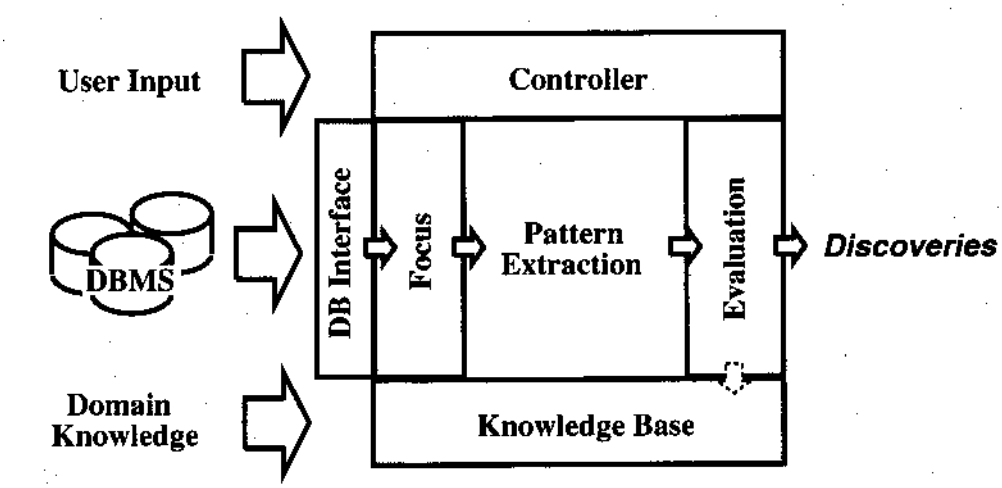
Spatiaalisen tietojen louhinnan sovellusalueita useita. Tunnetuin sovellusalue lienee maantieteessä käytetyt ns. GIS-järjestelmät (Geographic Information Systems). Myös esimerkiksi monet biotieteelliset tutkimusalueet, kuten molekyylibiologia, genetiikka ja aivotutkimus tuottavat runsaasti spatiaalista tietoa. Spatiaalista tietoa kertyykin jatkuvasti valtavia määriä. Esim. NASA:n järjestelmät tallentavat spatiaalista tietoa useita gigatavuja tunnissa. Koska tällaiset tietomassat eivät ole enää hallittavissa perinteisin menetelmin, tarve spatiaalisen tiedon automaattiselle käsittelylle ja analysoinnille kasvaa vastaavasti.

Tietojen louhinta spatiaalisista tietokantajärjestelmistä on monimutkaisempaa ja laskennallisesti raskaampaa kuin tietojen louhinta perinteisistä relaatiotietokannoista. Tämä on seurausta mm. käsiteltävän geometrisen tiedon monimutkaisuudesta sekä siitä, että louhinta-algoritmeissa pitää yleensä ottaa huomioon myös spatiaalisia tieto-olioita (spatial data objects) ympäröivien naapuriolioiden ominaisuudet. Tästä johtuen tietojen louhintaan sopivia tehokkaita järjestelmiä ei voida toteuttaa pelkästään perinteisten relaatiotietokantojen tarjoamia ominaisuuksia käyttäen, vaan lisäksi on hyödynnettävä erityisiä spatiaalisen tiedon käsittelyyn suunniteltuja tietotyyppejä, tietorakenteita ja algoritmeja.

Tässä esitelmässä pyrin luomaan yleiskuvan spatiaalisessa tietojen louhinnassa käytettävien tietokantajärjestelmien rakenteesta ja toiminnasta. Aluksi esittelen spatiaaliseen tietojen louhintaan sopivan järjestelmäarkkitehtuurin yleispiirteitä. Tämän jälkeen käsittelen spatiaalisen tiedon tallennukseen käytettyjä tietokantaratkaisuja sekä spatiaalisen tiedon indeksointimenetelmiä ja indeksoinnissa käytettyjä tietorakenteita. Lisäksi käsittelen spatiaalisessa tiedon louhinnassa käytettyjä naapurustokaavioita (neighbourhood graphs) ja naapurustoindeksejä (neighbourhood indeces). Lopuksi esittelen lyhyesti GeoMiner-nimisen geografisen tiedon louhintaan kehitetyn järjestelmän.

2 Tiedonlouhintajärjestelmän arkkitehtuuri

Tiedonlouhintajärjestelmällä tarkoitetaan tässä mahdollisimman pitkälle automatisoitua järjestelmää, joka mahdollistaa kiinnostavan ja aikaisemmin tuntemattoman tietämyksen löytämisen suuresta tietomassasta. Yleisesti tiedonlouhintajärjestelmä voidaan käsittää koostuvaksi useasta ohjelmistokomponentista, joista jokainen hoitaa tiettyä tiedonlouhinnan osatehtävää [MCP93]. Kuvassa 1 on kaavio tällaisesta monikomponenttijärjestelmästä, joka sisältää seuraavat peruskomponentit: ohjain (controller), tietokantaliittymä (database interface), tietämyskanta (knowledge base), kohdistus (focus), kuvioiden etsintä (pattern extraction) ja arviointi (evaluation). Tämä jaottelu on käsitteellinen, ja sen tarkoitus on luoda yksinkertainen kuva tiedonlouhintaan käytettyjen järjestelmien yleisrakenteesta. Käytännön toteutuksessa ole välttämättä mahdollista erottaa kaikkia tässä esitellyn mallin sisältämiä komponentteja.



Kuva 1. Tiedonlouhintajärjestelmän arkkitehtuuri [MCP93]

Kuvassa 1 esitetyssä järjestelmässä ohjainkomponentti ottaa vastaan käyttäjän komennot ja ohjaa näiden perusteella muiden komponenttien toimintaa tiedonlouhintaprosessin läpiviemiseksi. Ideaalitilanteessa ohjainkomponentti suorittaisi koko tiedonlouhintaprosessin täysin autonomisesti. Nykyisissä järjestelmissä ohjainkomponentti ei kuitenkaan ole autonominen, vaan käyttäjä joutuu yleensä osallistumaan tiedonlouhinnassa vaadittujen päätösten tekoon.

Kohdistuskomponentin avulla järjestelmä valitsee analyysin kohteeksi vain osan kaikista tietokannan sisältämistä tiedoista. Tiedonhakua voidaan rajoittaa esim. tiettyihin tauluihin ja tiettyihin tietokenttiin. Kohdistuskomponentti voi myös tarvittaessa ottaa satunnaisotoksia suurista tietomassoista analyysin nopeuttamiseksi.

Tietämyskanta sisältää tiedonlouhinnan tehostamisessa käytettävää aihepiirikohtaista tietoa (domain knowledge), joka voi koostua esimerkiksi rajoitemäärittelyistä tai alan asiantuntijoiden tuottamista käsitehierarkioista (concept hierarchies).

Spatiaalisen tiedon interaktiivinen käsittely vaatii yleensä sekä tulosten graafisen esittämisen käyttäjälle että mahdollisuuden syöttää kyselyiden parametreja graafisessa muodossa [Güt94]. Perinteisiin ei-spatiaalisiin kyselyihin liitetyt vakiot yleensä tekstimuotoista tietoa esim. ”kaupunki = ’Helsinki’”. Spatiaalisten kyselyiden tapauksessa käyttäjän pitää voida valita kohteita esim. kartalta tai edellisen kyselyn graafisesti esitetyistä tuloksista (esim. kyselyn rajoittaminen edellisen kyselyn tuloksena saadulle alueelle).

Varsinainen käsiteltävä tieto noudetaan tietokantajärjestelmästä tietokantaliittymässä luotujen kyselyjen (queries) avulla. Kyselyiden avulla noudettu tieto viedään analyysikomponentille, joka sisältää varsinaiset louhinta-algoritmit. Arviointikomponentin avulla järjestelmä valitsee esim. löydetyistä assosiaatiosäännöistä oleellisimmilta vaikuttavat säännöt, jotka järjestelmä sitten palauttaa käyttäjälle. Saadut tulokset voidaan lisäksi viedä tietämyskantaan käytettäväksi seuraavilla analyysikerroilla.

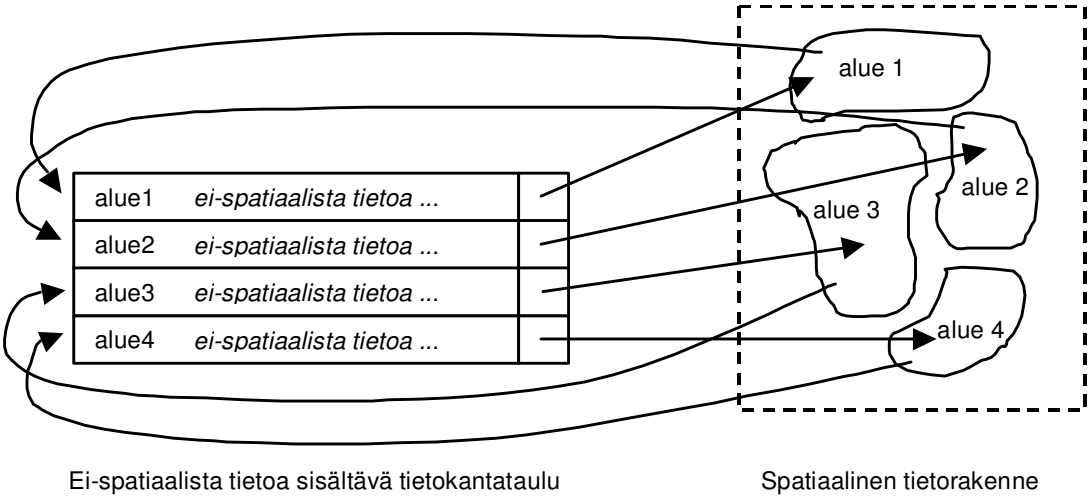
Tietokantaliittymän toteutus ja sen liittäminen tietokantajärjestelmään on keskeisessä asemassa tehokkaan tiedonlouhinnan kannalta [EKX95]. Seuraavassa luvussa esitellään spatiaalisten tietokantajärjestelmien erityispiirteitä verrattuna tavallisiin relaatiotietokantoihin ja kuvataan, miten spatiaalinen tietokantajärjestelmä voidaan toteuttaa hyödyntäen olemassa olevia tietokantajärjestelmiä.

Spatiaaliset tietokantajärjestelmät

Spatiaalisiin tieto-olioihin liittyy yleensä sekä spatiaalista että ei-spatiaalista tietoa (non-spatial data). Spatiaaliset tietokantajärjestelmillä (spatial database systems, SDBS) tarkoitetaan tietokantajärjestelmiä, joiden avulla voidaan käsitellä tavallisen ei-spatiaalisen tiedon lisäksi suuria geometrisista elementeistä koostuvia vektorimuotoisia tietomassoja (vector data) tai rasterimuotoisesta kuvatiedosta koostuvia tietomassoja (raster data) [Güt94]. Seuraavassa keskitytään lähinnä vektorimuotoisen datan käsittelyssä käytettäviin menetelmiin ja tietorakenteisiin.

2.1 Spatiaalisen tiedon mallinnus

Spatiaalisessa tietokannassa ei-spatiaalinen tieto sijaitsee tavallisissa tietokantarelaatioissa ja spatiaalinen tieto spatiaalisissa tietorakenteissa [AS91]. Spatiaalisten ja ei-spatiaalisten tietojen välillä ylläpidetään kaksisuuntaista viittauksia (kuva 2). Ei-spatiaalisten relaatioiden sisältämistä monikoista viitataan vastaaviin spatiaalisiin alkioihin spatiaalisessa tietorakenteessa ja näistä alkioista puolestaan viitataan takaisin vastaaviin ei-spatiaalisiin monikoihin.



Kuva 2. Spatiaalinen tietokanta [AS91]

Ei-spatiaalisen tiedon käsittelyyn on jo olemassa monia kehittyneitä tietokantajärjestelmiä. Sopivimmaksi tavaksi toteuttaa spatiaalinen tietokantajärjestelmä käytännössä onkin ehdotettu spatiaalisten tietorakenteiden ja operaatioiden liittämistä olemassa olevaan ns. laajennettavaan (extensible) tietokannanhallintajärjestelmään, esim. Postgres-järjestelmään [Güt94]. Tällaista ratkaisua kutsutaan integroiduksi arkkitehtuuriksi (integrated architecture).

Jotta erilaiset spatiaalisessa tietokantajärjestelmässä tarvittavat perusoperaatiot olisi mahdollista toteuttaa, pitää käsiteltävä spatiaalinen tieto mallintaa ns. spatiaalisilla tietotyypeillä (spatial data types) [Güt94]. Näitä ovat esim. piste (point), viiva (line) ja alue (region). Spatiaalisten tietotyyppien organisointiin voidaan käyttää erilaisia spatiaalisia tietorakenteita ja niihin kohdistuvia algoritmeja. Nämä mahdollistavat spatiaalisten tieto-olioiden ja operaatioiden toteuttamisen sekä liittämisen olemassa olevan tietokantajärjestelmän kyselyarkkitehtuuriin, esim. SQL-kielen laajennuksina..

Kun spatiaaliset tietorakenteet ja operaatiot toteutetaan valmiin tietokannanhallintajärjestelmän päälle, tarjoaa pohjana toimiva järjestelmä tiedonlouhintajärjestelmälle tietokannanhallinnan perusominaisuudet, kuten tehokkaan tiedon varastoinnin, eheyden hallinnan ja valmiita indeksointirakenteita [EKS01]. Tällainen järjestelmä käsittelee tavallisia tietotyyppejä ja spatiaalisia tietotyyppejä periaatteessa samalla tavoin. Vastaavasti myös tarvittavat indeksit, valinta- ja liitosoperaatiot sekä kyselyn optimointi käyttäytyvät periaatteessa samoin sekä ei-spatiaalisilla että spatiaalisilla tietotyypeillä [Güt94].

2.2 Spatiaalisen tiedon indeksointi

Spatiaalisen tiedon indeksoinnin (spatial indexing) tarkoitus on tukea spatiaaliseen tietoon kohdistuvia kyselyoperaatioita, esim. spatiaalista valintaa (spatial selection) [Güt94]. Spatiaalinen indeksi organisoi tilan ja siinä olevat kohteet siten, että kyselyoperaatioiden tarvitsee käydä läpi ainoastaan pieni osa spatiaalisten kohteiden koko joukosta. Ideana on välttää mahdollisimman paljon kyselyoptimoinnin kannalta verraten kallista spatiaalisten kohteiden tarkkojen geometristen ominaisuuksien vertailua.

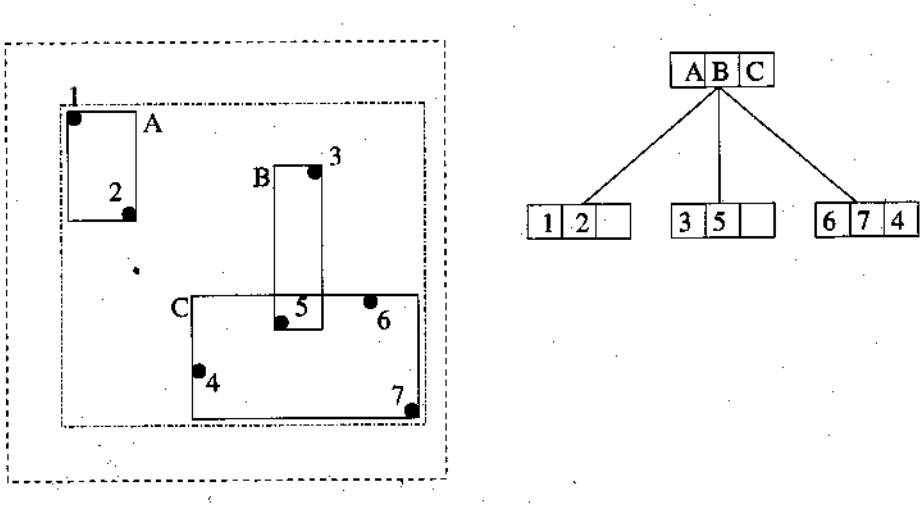
Spatiaalisessa indeksoinnissa käytetään hyväksi geometrisia approksimaatioita [Güt94]. Indeksirakenteeseen tallennetaan spatiaalisia kohteita vastaavia spatiaalisia avainarvoja

(spatial keys), jotka ovat todellisia spatiaalisia tietotyyppejä paljon yksinkertaisempia, esim. suorakaiteita, ja näin ollen myös huomattavasti nopeampia käsitellä.

Approksimaatioiden käytöstä on seurauksena ns. “suodata ja tarkenna” -kyselystrategia (filter and refine strategy) [Güt94]. Suodatusvaiheessa etsitään approksimaatioiden avulla joukko spatiaalisia kohteita, jotka suunnilleen täyttävät kyselyehdon. Tarkennusvaiheessa tähän joukkoon kuuluvien kohteiden tarkkaa geometriaa verrataan kyselyehtoon.

2.3 R-puut

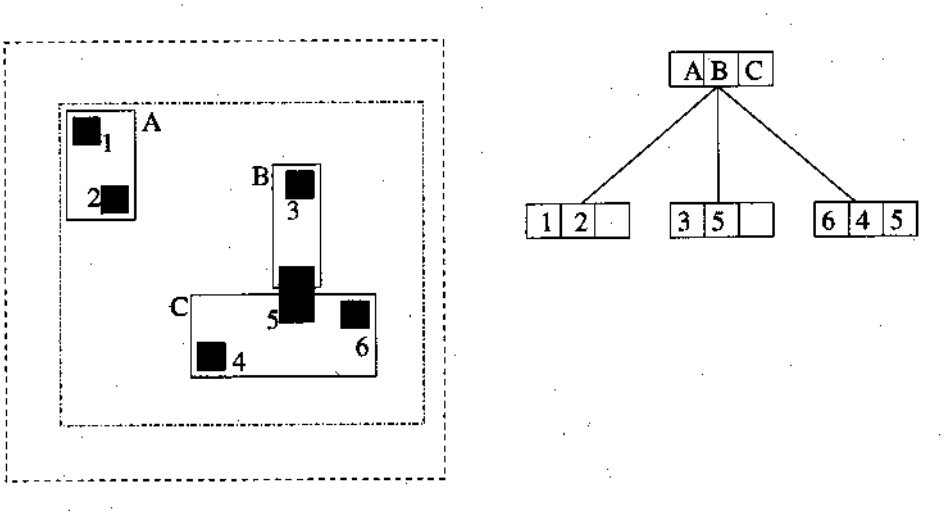
Tavalliset tietotyypit indeksoidaan yleensä B-puiden avulla, spatiaalisten tietotyyppien indeksoinnissa voidaan käyttää R-puuta tai sen muunnelmia [Kub01]. R-puu (kuva 3) on spatiaalisen tiedon indeksointiin tarkoitettu muunnelma B-puusta. R-puu on rakenteeltaan tasapainotettu puu ja jakaa tilan suorakaiteisiin, jotka voivat olla päällekkäisiä. R-puun jokainen solmu on ns. rajaava suorakaide (bounding box), joka sisältää kaikki kyseisen solmun alapuun solmut. Jokainen R-puun lehtisolmu sisältää osoittimen varsinaiseen spatiaaliseen tieto-olioon. Koska rajaavat suorakaiteet voivat olla päällekkäisiä, kyselyissä voidaan joutua käymään läpi useampia alapuita. Tästä johtuen rajaavat suorakaiteet pyritään pitämään toisistaan erillisinä. Lisättäessä puuhun uusia solmuja, pyritään INSERT-operaatiossa minimoimaan puuhun kohdistuvat rakenteelliset muutokset.



Kuva 3. R-puu [Kub01]

R*-puu on tavallisen R-puun muunnelma, jossa pyritään vähentämään turhia alapuiden läpikäyntejä minimoimalla kunkin rajaavan suorakaiteen sisään jäävä ”kuollut” tila, ts. tila joka ei kuulu millekään alapuun haaralle [Kub01]. R*-puut ovat tehokkaampia, mutta vaikeampia toteuttaa, kuin R-puut.

R+-puu (kuva 4) on R-puun laajennus, jossa rajaavat suorakaiteet eivät ole koskaan päällekkäin [Kub01]. Tämä rajoittaa tehokkaasti läpikäytävien alapuiden määrää. Puun rakenteen muutoksen seurauksena päällekkäiset rajaavat suorakaiteet jaetaan useammaksi pienemmäksi suorakaiteeksi, jotka voivat sisältää samoja kohteita. R+-puussa yksi spatiaalinen tieto-olio voi siis kuulua useaan samantasoiseen solmuun. R+-puu vaatii tavallista R-puuta enemmän tilaa, mutta vähentää kyselyissä indeksin läpikäyntiin tarvittavaa aikaa.



Kuva 4. R+-puu [Kub01]

On osoitettu, että esim. spatiaalisia klusterointioperaatioita voidaan huomattavasti tehostaa käyttämällä R*-puuta tiedonlouhintajärjestelmän tietokantaliittymäkomponentissa [EKX95]. R*-puun avulla voidaan suuresta tietokannasta poimia tehokkaasti laadukas otos spatiaalisia kohteita valitsemalla otokseen jokaisen rajaavan suorakaiteen (bounding box) sisältämistä kohteista keskimmäisin. Varsinainen klusterointialgoritmi kohdistetaan saatuun otokseen.

3 Tiedon louhinta spatiaalisesta tietokantajärjestelmästä

Spatiaalisissa tietojenlouhinta-algoritmeissa joudutaan yksittäisten käsiteltävien tieto-olioiden lisäksi tarkastelemaan näiden olioiden lähistöllä sijaitsevia spatiaalisia tieto-olioita. Tällöin tarkastellaan eri tieto-olioiden välisiä spatiaalisia suhteita kuvaavia ns. naapurustorelaatiota (neighbourhood relations) [EKS01]. Naapurustorelaatiot voidaan jakaa kolmeen perustyyppiin: topologisiin relaatioihin (topological relations), etäisyysrelaatioihin (distance relations) ja suuntarelaatioihin (direction relations).

Topologiset relaatiot kuvaavat kahden spatiaalisen kohteen päällekkäisyyksiä, erotuksia ja sisäkkäisyyksiä, esim. ”A on B:n sisällä”. Etäisyysrelaatioilla voidaan verrata kahteen spatiaalisen kohteen välistä etäisyyttä johonkin vakioon, esim. ”A on 5 km päässä B:stä”. Suuntarelaatiot puolestaan kuvaavat spatiaalisen kohteen suuntaa suhteessa toiseen kohteeseen, esim. ”A on B:stä itään”. Yhdistelemällä näitä perusrelaatioita loogisilla operaattoreilla voidaan määritellä monimutkaisempia spatiaalisia relaatioita, esim. ”A on B:stä itään ja korkeintaan 5 km päässä B:stä”.

3.1 Naapurustokaaviot, -polut ja -indeksit

Spatiaalisesta relaatiosta voidaan muodostaa ns. spatiaalisia naapurustokaavioita (spatial neighbourhood graphs) [EKS01]. Naapurustokaavio voi sisältää useita ns. naapurustopolkuja (neighbourhood paths). Naapurustokaavio ja -polku on määritelty seuraavasti:

> Olkoon *naapuri* naapurustorelaatio ja *DB* spatiaalisia kohteita sisältävä tietokanta. *Naapurustokaavio* $G^{DB}_{naapuri} = (N, E)$ on kaavio, joka sisältää solmut $N = DB$ ja reunat $E \subseteq N \times N$, jossa reuna $e = (n_1, n_2)$ on olemassa, jos ja vain jos relaatio *naapuri*(n_1, n_2) on tosi. *Naapurustopolku*, jonka pituus on k, on sarja solmuja $[n_1, n_2, ..., n_k]$, jossa *naapuri*(n_i, n_{i+1}) on tosi kaikille $n_i \in N$, $1 \leq i < k$.

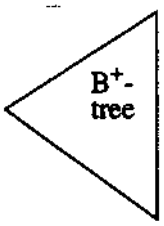
Naapurustokaavioille ja -poluille voidaan määrittää spatiaalista louhintaa tukevia operaatioita, esim. *naapurit* (neighbours) [EKS01]. *Naapurit (kaavio, kohde, ehdot)* palauttaa joukon spatiaalisia kohteita, jotka kytkeytyvät annettuun kohteeseen annetussa kaaviossa, ja jotka täyttävät annetut ehdot. Ehdot voivat perustua sekä spatiaalisiin että

ei-spatiaalisiin attribuutteihin, ja niiden tehtävänä on rajoittaa käsiteltävien naapurustopolkujen lukumäärää.

Spatiaalisten naapurien etsiminen tietokannasta on suhteellisen hidas operaatio johtuen spatiaalisten kohteiden monimutkaisuudesta. Lisäksi joudutaan yleensä suorittamaan useita *naapurit*-operaatioita. Toisaalta spatiaalisten tietokantojen sisältämä tieto on yleensä melko muuttumatonta. Em. syistä johtuen on ehdotettu, että varsinaisten spatiaalisten kohteiden käsittelemisen sijasta käsiteltäisiin naapurustokaavioita, jotka on tallennettu erilliseen naapurustoindeksiin (neighbourhood index) [EKS01].

Jokainen naapurustoindeksin rivi sisältää spatiaalisen olioparin sekä niiden välisen relaation (kuva 4). Naapurustoindeksissä sijaitsevien olioparien määrää voidaan rajoittaa asettamalla maksimietäisyys, joka parin välillä voi olla. Yksi naapurustoindeksi voi sisältää joukon erilaisia naapurustokaavioita. Naapurustoindeksi on määritelty formaalisti [EKS01]:

> Olkoon *DB* joukko spatiaalisia kohteita ja *max* ja *dist* reaalilukuja. Olkoon O_1 ja O_2 spatiaalisen joukon *DB* pisteitä, *D* suuntarelaatio ja *T* topologinen relaatio. Tällöin *naapurustoindeksi DB*:lle maksimietäisyydellä *max*:
> $I^{DB}_{max} = \{(O_1,O_2,dist,D,T) | O_1, O_2 \in DB, O_1 etäisyys_{=dist} O_2 \wedge dist \leq max \wedge O_2 D O_1 \wedge O_2 T O_1\}$



Object-ID	Neighbour	Distance	Direction	Topology
o_1	o_2	2.7	southwest	disjoint
o_1	o_3	0	northwest	overlap
...	...	...	...	...

Kuva 5. Naapurustoindeksi [EKS01]

Edellä kuvattu naapurustoindeksi voidaan luoda suorittamalla tietokantaan spatiaalinen liitosoperaatio (spatial join) relaation ($O_1 etäisyys_{=dist} O_2 \wedge dist \leq max$) perusteella [EKS01]. Operaatio voidaan suorittaa käyttämällä apuna edellä kuvattua spatiaalista indeksiä (ks. luku 2.3.). Liitosoperaation tuloksena saadut rivit tallennetaan tauluun, joka voidaan indeksoida attribuutin O_1 perusteella esim. B-puuhun. Naapurustoindeksi tehostaa esim.

naapurit-operaation suoritusta huomattavasti. Erityisen tehokas rakenne se on käsiteltäessä monimutkaisia spatiaalisia kohteita, esim. GIS-sovelluksissa. *Naapurit*-algoritmi voitaisiin toteuttaa naapurustoindeksiä hyödyntäen kuvassa 6 esitetyllä tavalla.

```
naapurit (Kaavio G, Olio O, Relaatio r, Ehto pred)
   hae muuttujaan Ind pienin sovellettava naapurustoindeksi kaaviossa G;
   if  (on olemassa Ind)
      etsi indeksistä Ind joukko oliota E, joille on indeksissä Ind
      merkintä (O,E,dist,D,T), ja sijoita ne muuttujaan ehdokkaat;
   else
      etsi tietokannasta spatiaalista indeksiä käyttäen relaation O r E
      toteuttava joukko olioita E ja sijoita ne muuttujaan ehdokkaat;
   naapurit = tyhjä joukko;
   for each E in ehdokkaat do
      if O r E and pred(E)
            lisää E joukkoon naapurit;
   return naapurit;
```

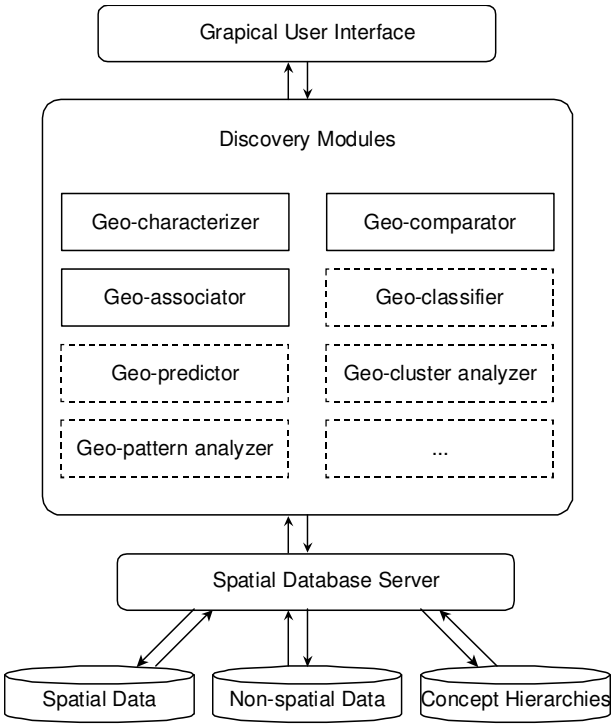
Kuva 6. Naapurit-algoritmi [EKS01]

4 GeoMiner

GeoMiner on DBMiner-nimisen relaatiotietokantapohjaisen tiedon louhintajärjestelmän laajennukseksi kehitetty monikomponenttiarkkitehtuuriin pohjautuva järjestelmä geo-spatiaalisen tiedon louhimiseen [JKS97]. GeoMiner käyttää spatiaalisen tiedon mallinnukseen edellä kuvattua järjestelyä, jossa käsiteltävä tieto on tallennettu toisiinsa linkitettyihin spatiaalisiin ja ei-spatiaalisiin tietorakenteisiin.

4.1 Yleisarkkitehtuuri

Kuvassa 7 on esitelty GeoMinerin yleisarkkitehtuuri. Varsinainen GeoMiner koostuu joukosta erilaisiin tiedon louhintatehtäviin tarkoitettuja moduuleja, mm. yleispiirteiden hahmottamiseen (Geo-characterizer), vertailusääntöjen louhimiseen (Geo-comparator) ja assosiaatiosääntöjen louhimiseen (Geo-associator). Moduuliarkkitehtuurin ansiosta GeoMinerin toiminnallisuutta voidaan tarvittaessa laajentaa toteuttamalla siihen uusia tiedonlouhintamoduuleita.



Kuva 7. GeoMiner-arkkitehtuuri [JKS97]

Varsinaisten louhintamoduulien lisäksi GeoMiner-arkkitehtuuriin kuuluu spatiaalinen tietokantajärjestelmä (esim. Informix-Illustra tai MapInfo), tietämyskanta (knowledge base) sekä vuorovaikutteisen tiedonlouhinnan ja tulosten graafisen esittämisen mahdollistava graafinen käyttöliittymä.

4.2 GMQL

GeoMineriin on toteutettu erityinen spatiaalisen tiedon louhintaan tarkoitettu SQL-kyselykielen laajennus *Geo-Mining Query Language* eli GMQL. Esimerkiksi Queenslandin eri lämpötila- ja kosteusalueiden hahmottaminen voitaisiin GMQL:ää käyttämällä suorittaa seuraavasti:

```
mine spatial characteristic
as "Queenslandin lämpötila- ja kosteusjakauma"
analyze geo
in relevance to temperature, precipation
from weather_probe
where time_period = "summer" and year = 2002
      and area_name = "Queensland"
```

Operaation suorittamiseksi järjestelmä noutaa ensin spatiaalisesta tietokannasta tarvittavan tietojoukon ja suorittaa tälle ei-spatiaalisiin tietoihin (lämpötila ja kosteus) perustuvan yleistyksen. Esim. lämpötilat voivat yleistyksessä jakautua käsiteluokkiin *viileä, lämmin, kuuma* jne. Tämän jälkeen järjestelmä yhdistää yleistysten perusteella samankaltaiset spatiaaliset alueet toisiinsa.

Käsitehierarkiat voivat olla käyttäjien määrittelemiä tai perustua klusterointialgoritmien tuloksiin. Edellä kuvatussa operaatiossa yleistysalgoritmi perustui lähtökohtaisesti ei-spatiaalisiin käsitehierarkioihin perustuvaa, mutta yleistys voi toisentyyppisissä kyselyissä perustua vastaavalla tavalla myös spatiaalisiin käsitehierarkioihin.

5 Yhteenveto

Spatiaalisella tiedolla tarkoitetaan tietoa, joka liittyy sijaintiin tai geometriseen muotoon. Spatiaaliset tietokantajärjestelmien avulla voidaan tehokkaasti käsitellä suurista määristä geometrisista elementeistä tai rasteripisteistä koostuvaa spatiaalista tietoa sekä siihen liittyvää ei-spatiaalista tietoa. Spatiaalisissa tietokantajärjestelmissä spatiaalinen tieto on yleensä mallinnettu erilaisten spatiaalisten tietotyyppien, kuten pisteiden ja viivojen, avulla. Yksinkertaisista spatiaalisista tietotyypeistä koostuva tietomassa puolestaan organisoidaan erilaisten spatiaalisten tietorakenteiden ja niihin kohdistuvien algoritmien avulla.

Spatiaalisen tiedon louhinnalla tarkoitetaan kiinnostavan ja yleensä aikaisemmin tuntemattoman tietämyksen löytämistä spatiaalista tietoa sisältävistä tietokannoista. Spatiaalinen tiedonlouhintajärjestelmä on automatisoidun tiedonlouhintaprosessin mahdollistava monikomponenttijärjestelmä. Eräs tällainen olemassa oleva tiedonlouhintajärjestelmä on geo-spatiaalisen tiedon louhintaan kehitetty GeoMiner.

Ideaalinen tiedonlouhintajärjestelmä suorittaisi koko tiedonlouhintaprosessin automaattisesti, mutta käytännön järjestelmät ovat tästä tavoitteesta toistaiseksi vielä varsin kaukana. Tulevaisuudessa olisikin erityisesti tarvetta monille eri sovellusalueille soveltuvalle hyvin pitkälle automatisoidulle tiedonlouhintajärjestelmälle.

Lähteet

AS91 Aref, W. G., Samet, H., *Optimization strategies for spatial query processing*. Proc. 17th international conference on very large databases (VLDB), Barcelona, Spain, 1991. 81-90.

EKS01 Ester, M., Kriegel, H., Sander, J., *Algorithms and applications for spatial data mining*. Geographic Data Mining and Knowledge Discovery, Research Monographs in Geographic Information Systems, Taylor and Francis, 2001.

EKX95 Ester, M., Kriegel, H., Xu, X., *Knowledge discovery in large spatial databases: focusing techniques for efficient class identification*. Proc. 4th international symposium on large spatial databases (SSD'95), Berlin, 1995. 67-82.

Güt94 Güting, R. H., *An introduction to spatial database systems*. VLDB Journal, Special issue on spatial database systems, 3(4), 1994. 357-399.

JKS97 Jiawei, H., Koperski, K., Stefanovic, N., *GeoMiner: a system prototype for spatial data mining*. Proc. ACM SIGMOD international conference on management of data, Arizona, 1997. 553-556. *http://db.cs.sfu.ca/GeoMiner/*

Kub01 Kuba, P., *Data structures for spatial data mining*. FI MU report series, Masaryk University, 2001.

MCP93 Matheus, C. J., Chan, P. K., Piatetsky-Shapiro, G., *Systems for knowledge discovery in databases*. IEEE transactions on knowledge and data engineering, 5(6), 1993. 903-913.