Spatiaalisten assosiaatiosääntöjen louhinta

Saija Lemmelä

Helsinki 18.04.2003
Spatiaalisen tiedon louhinta -seminaari
HELSINGIN YLIOPISTO
Tietojenkäsittelytieteen laitos

SISÄLLYS

1. JOHDANTO 3

2. SPATIAALISET OBJEKTIT, PREDIKAATIT JA ASSOSIAATIOSÄÄNNÖT 3

2.1 Spatiaalinen assosiaatiosääntö 4

3. SPATIAALISTEN ASSOSIAATIOSÄÄNTÖJEN LOUHINTA 5

3.1 Koperskin ja Hanin menetelmä 6

3.2 Tiedon esitystapa ja taustatietojen esittäminen 7
3.2.1 Esimerkki spatiaalisten assosiaatiosääntöjen louhinnasta 8

3.3 Epävarman tiedon louhinta 9
3.3.1 Alueet ja reunusalueet 10
3.3.2 Topologisten riippuvuuksien käsitehierarkia 11
3.3.3 Päätöspuun käyttö predikaattien arvojen tutkinnassa 13
3.3.4 Esimerkki käsitehierarkiaa hyödyntävästä louhinta-algoritmista 15

4. YHTEENVETO 15

1. JOHDANTO

Spatiaalisen tiedon louhinta eli mielenkiintoisen, implisiittisen tiedon etsintä spatiaalisesta tietokannasta, on tärkeä väline hyödynnettäessä spatiaalisten tietokantojen tietosisältöä. Tässä työssä tarkastellaan erityisesti spatiaalisten assosiaatiosääntöjen louhintaa. Spatiaaliset assosiaatiosäännöt kuvaavat spatiaalisessa tietokannassa usein esiintyviä spatiaalisten objektien keskinäisiä tai spatiaalisten ja ei-spatiaalisten objektien välisiä riippuvuuksia.

Koska spatiaaliset tietokannat ovat yleensä suuria, pyritään spatiaalisia assosiaatiosääntöjä louhittaessa yleensä mahdollisimman aikaisessa vaiheessa rajaamaan laskenta kulloisenkin tehtävän kannalta oleellisiin objekteihin. Taustatietojen ja käsitehierarkioiden käyttäminen spatiaalista tietoa louhittaessa on varsin yleistä.

Spatiaalinen tieto kuvaa usein mitattua luonnonilmiöitä kuvaavaa tietoa. Tämän vuoksi tieto on usein epätäydellistä, ristiriitaista, epämääräistä, epätarkkaa tai virheellistä. Kun spatiaaliseen tietoon liittyy epävarmuutta, louhintatekniikan tulee pystyä suodattamaan epävarmuustekijät pois ilman että tietoon tarvitsee tehdä karkeita yksinkertaistuksia

Tässä työssä esitellään joitakin spatiaalisten assosiaatiosääntöjen louhinnassa käytettyjä menetelmiä. Luvussa 2 esitellään assosiaatiosääntöihin liittyvää käsitteistöä. Luku 3 keskittyy louhintamenetelmien esittelyyn. Luvussa esitellään Koperskin ja Hanin käyttämä louhinta-algoritmi sekä esitellään tapoja, joilla taustatietoja, erityisesti käsitehierarkioita voidaan hyödyntää spatiaalisen tiedon louhintaprosessissa. Luku 4 sisältää yhteenvedon.

2. SPATIAALISET OBJEKTIT, PREDIKAATIT JA ASSOSIAATIOSÄÄNNÖT

Spatiaalisten assosiaatiosääntöjen louhinta on assosiaatioiden etsimistä viiteobjektien (reference objects) ja tehtävän kannalta relevanttien objektien (task relevant objects) väliltä. Viiteobjekti voi olla esimerkiksi suuri kaupunki, jonka läheisyydessä olevia teitä ja vesistöjä (tehtävän kannalta relevantit objektit) halutaan tutkia.

Verrattaessa spatiaalisten assosiaatiosääntöjen louhintaa relaatiotietokannasta suoritettavaan assosiaatiosääntöjen louhintaan merkittävämmäksi eroksi nousee:

- spatiaalisten riippuvuuksien implisiittisyys ja
- spatiaalisten objektien monitasoisuus [MEL02].

Näistä ensimmäinen johtuu spatiaalisten objektien sijainnista ja laajuudesta, jotka kuvaavat spatiaaliset suhteet implisiittisesti. Spatiaalisten suhteiden muuntamiseksi eksplisiittisiksi tarvitaan yleensä monimutkainen muunnosprosessi. Objektien monitasoisuus taas on seurausta spatiaalisten objektien mahdollisesta sisäkkäisestä sijainnista.

Spatiaaliset riippuvuudet (spatiaaliset predikaatit) voidaan jakaa kolmeen aliryhmään:

- topologisiin riippuvuuksiin spatiaalisten objektien välillä, kuten *sisällä, ulkopuolella, naapurissa, leikkaa, peittää,*
- sijaintiin tai järjestykseen liittyviin riippuvuuksiin, kuten *oikealla, vasemmalla, pohjoisessa,* ja
- etäisyyttä kuvaavaan tietoon, kuten *lähellä* ja *kaukana.*

Monitasoisen tiedon esittämiseen käytetään yleisesti käsitehierarkioita. Käsitehierarkiassa tieto muuttuu yleisluontoisemmaksi hierarkiassa ylöspäin mentäessä, mutta säilyy silti yhdenmukaisena alemmilla käsitetasoilla olevan tiedon kanssa. Esimerkkinä käsitehierarkiasta voidaan esimerkiksi ajatella valtion jakautumista lääneihin ja läänien edelleen kuntiin. Yleisesti käytetyt käsitehierarkiat ovat usein asiantuntijoiden luomia tai automaattisesti generoituja. Käsitehierarkia voi myös olla koodattu tietokannan skeemaan.

2.1 Spatiaalinen assosiaatiosääntö

Spatiaaliset assosiaatiosäännöt kuvaavat spatiaalisessa tietokannassa usein esiintyviä malleja. Säännöt kuvaavat joko spatiaalisten predikaattien keskinäisiä assosiaatiosuhteita tai niiden ja ei-spatiaalisten predikaattien välisiä suhteita. Spatiaalinen assosiaatiosääntö voidaan määritellä seuraavasti:

$$X_1 \wedge ... \wedge X_n \rightarrow Y_1 \wedge ... \wedge Y_n (c\%)$$

Tässä ainakin yksi $X_1 ... X_n, Y_1 ... Y_n$ on spatiaalinen predikaatti ja c% on säännön varmuus (confidence). Esimerkiksi sääntö ’80% kouluista sijaitsee lähellä puistoa’ yhdistää ei-spatiaalisen predikaatin *on* (is_a) spatiaaliseen predikaattiin *lähellä* (close_to). Tämä sääntö voidaan merkitä:

$$on(X, koulu) \rightarrow lähellä(X, puisto)\ (80\%)$$

Jotta spatiaalinen sääntö on tutkimisen kannalta merkityksellinen sen ilmaiseman mallin on esiinnyttävä suhteellisen säännöllisesti tietokannassa. Vahvaksi (strong) sääntö määritellään silloin, kun se sekä havainnollistaa riittävän suuren objektijoukon piirrettä (minimum support treshold) että sen varmuus on riittävän suuri (minumum confidence treshold). Esimerkiksi mikäli säännön varmuutta kuvaamaan on asetettu alaraja 90% sääntö:

$$on(X, kaupunki) \wedge sijaitsee(C, BritishColumbia) \wedge rannalla(X, vesistö) \rightarrow lähellä(X, USA)\ (92\%)$$

täyttää varmuuskriteeteerin. Koska sääntö ilmaisee kaikkien British Columbian kaupunkien piirrettä, voidaan myös havaintojen määrän katsoa olevan riittävä. Sen sijaan, mikäli sääntö rajoitettaisiin koskemaan pelkästään Victorian kaupunkia, voitaisiin sen esittämistä pitää turhana:

$$on(X, Victoria) \wedge sijaitsee(C, BritishColumbia) \wedge rannalla(X, vesistö) \rightarrow lähellä(X, USA)\ (100\%)$$

Vaikka riittävän suureen havaintojen määrään perustuvia sääntöjä, joiden varmuus on 100%, ei juurikaan ole, sisältävät spatiaaliset assosiaatiosäännöt usein kiinnostavaa epätriviaalia tietoa, joka on hyödyllistä tutkitettaessa muun muassa maantieteeseen, ympäristöön, biologiaan tai teknisiin ratkaisuihin liittyviä ilmiöitä.

3. SPATIAALISTEN ASSOSIAATIOSÄÄNTÖJEN LOUHINTA

Spatiaalisten assosiaatiosääntöjen louhinta on assosiaatioiden etsimistä viiteobjektien ja tehtävän kannalta relevanttien objektien väliltä. Louhittaessa spatiaalisia assosiaatiosääntöjä tehdään yleensä vaativammat laskentaoperaatiot suuremmalla abstraktiotasolla, jonka jälkeen keskitetään laskenta objekteihin, jotka näin saadun karkean erottelun perusteella vaikuttavat mielenkiintoisimmilta. Luvussa 3.1 esitetään tätä perusperiaatetta noudattava louhintamenetelmä. Tärkeässä asemassa spatiaalisia assosiaatiosääntöjä louhittaessa ovat myös käsitehierarkiat ja muu kohdealueeseen liittyvä taustatieto, jotka ovat hyödynnettävissä laskentaa tehtäessä. Tiedon esitystapaan ja taustatietoihin perehdytään luvussa 3.2. Luku 3.3 keskittyy spatiaaliseen tietoon liittyvien epävarmuustekijöiden eliminoinnin mahdollisuuksiin laajan reunusalueen käsitteen avulla.

3.1 Koperskin ja Hanin menetelmä

Yleisesti spatiaalisten assosiaatiosääntöjen louhinnassa käytetyn Koperskin ja Hanin [KoH95] [HKS97] esittelemän algoritmin perusajatuksena on ensin pienentää laskentaan mukaan otettavien tietokantaobjektien määrää jättämällä mahdollisimman aikaisessa vaiheessa laskennasta pois tuloksen kannalta epärelevantit objektit.

Algoritmi koostuu viidestä askelesta, jotka ovat pääpiirteissään seuraavat:

1. Etsitään tehtävän kannalta relevantit objektit paikkatietokannasta (SDB) ja relaatiotietokannasta (RDB), joka sisältää spatiaalisiin objekteihin liittyvää ei-spatiaalista tietoa. Tulos tallennetaan *Task_relevant_DB*- tietokantaan.

   Esimerkkitapauksena käsittelemme British Colombiassa sijaitsevia kaupunkeja, teitä, vesistöjä, kaivoksia ja USAn rajaa.

2. Suoritetaan *Task_relevant_DB*- kannassa jokin tehokas spatiaalinen laskenta-algoritmi (esimerkiksi R-tree [BKS93] tai MBR-tekniikka ja plane-sweep algoritmi [PrS95]) karkean erottelun aikaansaamiseksi. Tarkoituksena on selvittää suhteellisen lähellä toisiaan sijaitsevat objektit. Objektien välisiä suhteita kuvaavat spatiaaliset predikaatit talletetaan laajennettuun *Coarse_predicate_DB* relaatiotietokantaan.

   Esimerkiksi selvitetään karkealla tasolla suurten kaupunkien spatiaalinen läheisyys (*lähellä*-predikaatti) muihin edellä määriteltyihin luokkiin eli teihin, vesistöihin, kaivoksiin ja USAn rajaan.

3. Lasketaan kuhunkin predikaattiin liittyvien havaintojen määrä *Coarse_predicate_DB*-kannassa ja suodatetaan pois ne esiintymät, joihin liittyvien havaintojen määrä alittaa asetetun havaintojen sallitun minimimäärän (minimum support treshold). Tulos tallennetaan *Large_Coarse_predicate_DB*- kantaa. Tässä yhteydessä on huomattava, että suuressa havaintojoukossa mielenkiintoinen predikaatti on yleensä löydettävissä myös muilta käsitetasoilta.

   Esimerkiksi:

   $on(X, suuri\_kaupunki) \rightarrow lähellä\ (X, vesistö)\ (80\%)$

   $on(X, suuri\_kaupunki) \wedge\ lähellä(X, meri) \rightarrow lähellä\ (X, us\_raja)\ (92\%)$

4. Suoritetaan *Large_Coarse_predicate_DB*- kannassa tehokas spatiaalinen laskenta-algoritmi [BKS93][PrS95] sen sisältämien objektien välisten suhteiden tarkaksi selvittämiseksi.

   Esimerkiksi tarkennetussa spatiaalisessa laskennassa jokainen *lähellä*-predikaatti voidaan korvata *leikkaa, vieressä, sisällä* tai *lähellä*-predikaatilla.

5. Lasketaan kaikkien predikaattiyhdistelmien havaintomäärät ylimmällä käsitetasolla, ja suodatetaan pois ne yhdistelmät, joihin liittyvien havaintojen määrä on alle asetetun alarajan.

   Esimerkiksi:

k	k-predikaattien joukko	lkm
1	*<vieressä, vesistö>*	*32*
1	*<leikkaa, valtatie>*	*29*
2	*<vieressä, vesistö> <leikkaa,valtatie>*	*25*
	jne.	

6. Lopuksi generoidaan assosiaatiosäännöt usein toistuvien predikaattiyhdistelmien perusteella. Tämä tehdään kaikilla käsitetasoilla, kunnes riittävään havaintomäärään perustuvia sääntöjä ei enää löydy.

   Esimerkiksi:

   $on(X, suuri\_kaupunki) \wedge leikkaa(X, tie) \rightarrow vieressä(X, vesistö)$ *(86%)*

   $on(X, suuri\_kaupunki) \wedge leikkaa(X, tie) \rightarrow vieressä(X, meri)$ *(79%)*

   Säännön varmuus on laskettu havaintojen määrän perusteella esimerkiksi: 25/29=86%.

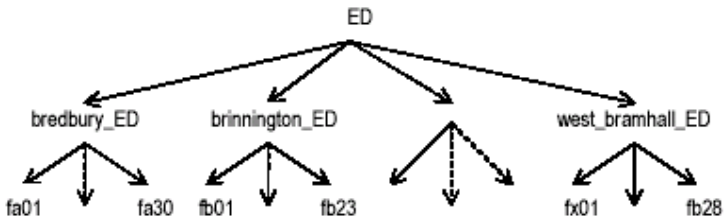
3.2 Tiedon esitystapa ja taustatietojen esittäminen

Malerba, Esposito ja Lisi [MEL01] [MEL02] esittävät että spatiaalinen tietokanta voidaan muuttaa deduktiiviseksi relaatiotietokannaksi, kun spatiaaliset suhteet viitebjektien ja tehtävän kannalta relevanttien objektien välillä on selvitetty. Tietokantojen kuvausvoima antaa myös mahdollisuuden kuvata sääntöjen avulla taustatietoja (background knownledge, BK), kuten spatiaalisia hierarkioita, ja tutkittavalle kohteelle ominaista tietoa.

SPADAssa (Spatial Pattern Discovery Algorithm) tieto esitetään Datalog-atomeina. Datalog on relaatiotietokannoissa käytettävä logiikkakieli joka sallii sekä spatiaalisten suhteiden, että symbolisten taustatietojen esittämisen. Tällaisia taustatietoja ovat esimerkiksi spatiaaliset hierarkiat, muuttujat ja säännöt, joita voidaan käyttää päättelyn apuna. SPADA pystyy käsittelemään esimerkiksi topologisia relaatioita. Itse louhinta-algoritmi SPADAssa on perinteinen tasoittain etenevä tiedonlouhinta-algoritmi [MaT97].

3.2.1 Esimerkki spatiaalisten assosiaatiosääntöjen louhinnasta

Esimerkkitapauksessa [MEL02] selvitettiin sosiologiselta näkökannalta moottoritie M63 palvelemaa aluetta siten, että tulokset olivat hyödynnettävissä liikennesuunnittelussa. Itse aluetta kuvataan tässä tapauksessa kahdella tarkkuustasolla muutaman assosiaatiosäännön avulla. Tutkittavat kunnat (ED) on ryhmitelty lääneittäin (ward) ja nämä on kuvattu Datalog faktoina. Alueiden hierarkia on esitetty kuvassa 1.



Kuva 1. Tutkittavien alueiden hierarkia

Spatiaaliset assosiaatiosäännöt kytkevät kunnat, joiden lävitse M63 kulkee (viitealueet) kuntiin, joita mootoritie M63 palvelee (tehtävän kannalta relevantit objektit). Leikkaussuhteet (kunta - moottoritie) ja naapuuruussuhteet (kunta - kunta) on selvitetty kiinnostuksen kohteena olevalta alueelta ja muutettu Datalog-faktoiksi. Tutkittaviksi on valittu erityisesti seuraavat attribuutit:

- Henkilöt, jotka asuvat yleisimmän asuinalueen ulkopuolella ja jotka ajavat töihin autolla.
- Työntekijät, jotka asuvat taloudessa, jossa on vähintään kolme autoa, ja jotka ajavat töihin autolla.
- Työntekijät, jotka asuvat taloudessa, jossa on vähintään kolme autoa, ja jotka työskentelevät yleisimmän asuinalueen ulkopuolella.

Jokainen kunta kuvataan D(S)ssä kolmella perusatomilla (ground atom): *dr_out(X, [a..b]), cars3_dir(X, [a..b])* ja *cars3_out(X, [a..b]).* Tässä X viittaa kuntaan ja [a..b] e.m. attribuutteihin, jotka on normalisoitu ja muunnettu numeerisista arvoista SPADA:n paremmin käsittelemään epäjatkuvaan muotoon. Viiteobjekteja S:ssä kuvaava avainatomi (key atom) *ed_on_M63(X*) on lisätty taustatiedoihin sääntönä:

ed_on_M63(X) :- intersect (X, m63)

Taustatietohin on myös kirjattu joitakin deklaratiivisia kuvauksia spatiaalisessa päättelyssä käytettävistä säännöistä:

can_reach(X,Y) :- intersect(X, m63), intersect (Y, m63), Y\=X

close_to(X,Y) :- adjanced_to(X, Z), adjanced_to(Z,Y), Y\=X

Lisäksi on asetettu alarajat liittyen sääntöjen esiintymistodennäköisyyteen ja havaintojen määrään: min_sup[1]=0.7 ja min_conf[1]=0.9 ensimmäisellä tasolla ja min_sup[2]=0.5 ja min_conf[2]=0.8 toisella tasolla.

Kun SPADA ajettiin painottaen attribuutteja *dr_out(X, [a..b]), cars3_dir(X, [a..b])* ja *cars3_out(X, [a..b])* saatiin 10531 vahvaa assosiaatiosääntöä esimerkiksi:

ed_on_M63(X), close_to(X, Y), is_a(Y,Bedgekey_ED) ->
cars3_out(X,[0.0…0.037]), cars_dr(Y, [0.0..0.037]) (100%, 100%)

Tämä voidaan lukea: ”Jos kunta (X), jota M63 leikkaa, on jonkin muun Bedgeleyn äänestysalueeseen kuuluvan kunnan (Y) lähellä, niiden kunnan (X) asukkaiden lukumäärä, joiden taloudessa on enemmän kuin 3 autoa ja jotka ajavat alueen ulkopuolelle töihin on hyvin pieni (alle 4% asukkaista).”

3.3 Epävarman tiedon louhinta

Käytössä olevat spatiaalisen tiedon louhintametodit eivät useinkaan huomioi spatiaaliseen tietoon liittyviä epävarmuustekijöitä. Edellä esitetty Koperskin ja Hanin algoritmi [KoH95] perustuu oletukseen, että spatiaalisten objektien sijainnit ovat tarkasti tiedossa. Spatiaalinen tieto on kuitenkin usein epätäydellistä, ristiriitaista, epämääräistä, epätarkkaa tai virheellistä. *Epätäydellistä* tieto on silloin kun se puuttuu joko kokonaan tai osittain. Tyypillisesti epätäydellisyyttä esiintyy esimerkiksi paperikartasta digitoidun viivan ollessa katkonaista.

Ristiriitaisuutta esiintyy silloin kun samasta objektista on useita keskenään erilaisia versioita, jotka on saatu esimerkiksi eri tietolähteistä tai jotka esiintyvät eri abstraktiotasoilla. Luonnonmaantieteelliset ominaisuudet ovat usein *epämääräisiä* johtuen siitä, että tyypillisesti luonnossa rajat eivät ole yleensä selkeitä ja yksiselitteisiä. *Epätarkkuus* johtuu spatiaalisten kokonaisuuksien esitystavasta. Tyypillinen esimerkki on rasteridata, jossa rasteriruutu on pienin tilaa kuvaava yksikkö. *Virhe* syntyy tyypillisesti esimerkiksi mittauksia tehtäessä [CFK00].

Clementini, Di Felice ja Koperski [CFK00] ovat esittäneet laajan reunusalueen (broad boundary) käyttöä silloin, kun spatiaalinen tieto on jossain suhteessa epävarmaa. Tämä lähestymistapa suodattaa pois spatiaaliseen tietoon usein liittyvät epävarmuustekijät ja mahdollistaa tietoon liittyvät laskentaoperaatiot ilman, että tietoon tarvitsee tehdä karkeita yksinkertaistuksia. Tätä tekniikkaa käytettäessä objektien väliset topologiset suhteet organisoidaan kolmitasoiseksi käsitehierarkiaksi.

Tätä käsitehierarkiaa käytetään louhintaprosessissa, joka aloitetaan korkeimmalta käsitetasolta. Korkeimmalta käsitetasolta löytyneiden usein toistuvien mallien osalta louhintaa jatketaan seuraavalla käsitetasolla. Louhintaa jatketaan näin taso tasolta alemmalle käsitetasolle, kunnes usein toistuvia malleja ei enää löydetä. Objektien välisten topologisten relaatioiden kuvaamiseen voidaan luoda päätöspuu (decision tree). Tämä puu pyritään luomaan sellaiseksi, että topologisten suhteiden määrittelemiseksi tarvittavien laskennallisten operaatioiden määrä on mahdollisimman pieni.

3.3.1 Alueet ja reunusalueet

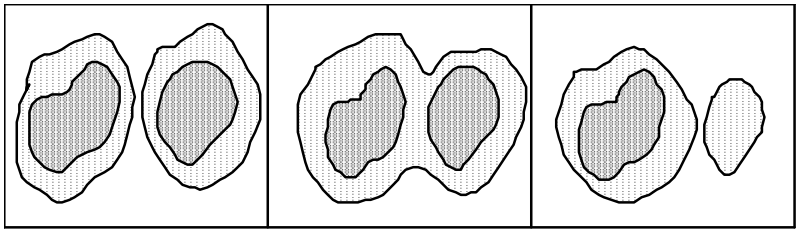
Alue (region) on säännöllinen suljettu kaksiulotteinen $\mathbb{R}^2$:n osajoukko, jolla on yhtenäinen sisusta. Alueessa voi siis olla reikiä. *Yhdistelmäalue* (composite region) on säännöllinen kaksiulotteinen $\mathbb{R}^2$:n osajoukko.

Alue, jolla on laaja reunusalue (A region with a broad boundary) A is koostuu kahdesta alueesta A_1 ja A_2, joissa $A_1 \subseteq A_2$. Tässä ∂A_1 on A:n sisempi reuna of A ja ∂A_2 sen ulompi reuna.

Laaja reunusalue (the broad boundary) ΔA on suljettu osajoukko sisemmän ja ulomman reunan välillä. Eli $\Delta A = \overline{A_2 - A_1}$, tai $\Delta A = A_2 - A_1^\circ$.

Laajan reunusalueen omaavan alueen sisus (interior), sulkeuma (closure) ja ulkopuoli (exterior) määritellään seuraavasti: $A^\circ = A_2 - \Delta A$, $\overline{A} = A^\circ \cup \Delta A$, $A^- = \mathbb{R}^2 - \overline{A}$. Laajan reunusalueen omaavan alueen sisä- ja ulkopuoli ovat avoimia joukkoja kun taas laaja reunusalue itse on suljettu joukko.

Yhdistelmäalueeseen liittyvä peruskäsitteistö määritellään vastaavalla tavalla kuin alueeseen liittyvä käsitteistö. Kuva 2 esittää joitakin mahdollisia yhdistelmäalueen kokoonpanoja: ensimmäisessä tapauksessa yhdistelmäalueella on kaksi komponenttia. Toisessa tapauksessa A_1 on kaksi komponenttia ja A_2 yksi. Kolmannessa tapauksessa A_1 on yksi komponentti ja A_2 kaksi komponenttia.



Kuva 2. Yhdistelmäalueita, joilla on laaja reunusalue

3.3.2 Topologisten riippuvuuksien käsitehierarkia

Clementini, Di Felice ja Koperski [CFK00] käyttävät topologisia suhteita kuvaavaa käsitehierarkiaa. Binäärinen topologinen suhde kahden objektin A:n ja B:n välillä $\mathbb{R}^2$:ssa voidaan luokitella A:n sisustan, reunan ja ulkopuolisen alueen ja B:n sisustan, reunan ja ulkopuolisen alueen välisen leikkauksen perusteella. Näiden kuuden objektin välillä voidaan muodostaa yhdeksän erilaista leikkausta, jotka määrittävät topologisen relaation. Tämä topologinen relaatio voidaan esittää seuraavana 3×3 matriisina *M*, jota kutsutaan nimellä *9-intersection*:

$$M = \begin{pmatrix} A^\circ \cap B^\circ & A^\circ \cap \Delta B & A^\circ \cap B^- \\ \Delta A \cap B^\circ & \Delta A \cap \Delta B & \Delta A \cap B^- \\ A^- \cap B^\circ & A^- \cap \Delta B & A^- \cap B^- \end{pmatrix}.$$

Käytettäessä arvoja tyhjä (0) ja epätyhjä (1) voidaan erotella 2^9=512 binääristä topologista relaatiota. Yksinkertaiselle alueille, jolla on yksiulotteinen reuna, vain kahdeksan näistä

relaatioista voi realisoitua. Kahdelle yhdistelmä-alueelle, joilla on yksiulotteinen reuna, on olemassa kahdeksan vaihtoehtoista matriisia ja 16 mahdollista erilaista relaatiota. Kun objekteilla on laajat reunusalueet mahdollisten relaatioiden määrä nousee 56:een. Yleisesti ottaen 9-intersection-metodi tarjoaa käyttäjälleen mahdollisuuden testata suurta spatiaalisten relaatioiden määrä ja tarkentaa tarkastelua testattavana olevaan relaatioon. Menetelmän heikkoutena on sen liika yksityiskohtaisuus monien käytännön sovellusten kannalta. Suurelle osalle alempien käsitetasojen suhteista ei myöskään löydy vastinetta luonnollisesta kielestä.

Laajan reunusalueen omaavien yhdistelmä-alueiden välisiä topologisia suhteita on tutkittu kolmella hierarkiatasolla. Alin hierarkiataso koostuu 56 relaatiosta, jotka määritellään 9-intersection matriisin kautta. Ylimmällä hierarkiatasolla on taas käytössä ainoastaan neljä relaatioita (*disjoint*, *touch*, *overlap* ja *in*). Ylimmän hierarkiatason topologiset suhteet on esitelty tarkemmin kaaviossa 1.

disjoint	$\begin{pmatrix} 0 & \delta & \delta \\ \delta & 0 & \delta \\ \delta & \delta & \delta \end{pmatrix}$
touch	$\begin{pmatrix} 0 & \delta & \delta \\ \delta & 1 & \delta \\ \delta & \delta & \delta \end{pmatrix}$
overlap	$\begin{pmatrix} 1 & \delta & \delta \\ \delta & \delta & 1 \\ \delta & 1 & \delta \end{pmatrix} \vee \begin{pmatrix} 1 & \delta & \delta \\ 1 & \delta & 1 \\ \delta & \delta & \delta \end{pmatrix} \vee \begin{pmatrix} 1 & 1 & \delta \\ \delta & \delta & \delta \\ \delta & 1 & \delta \end{pmatrix} \vee \begin{pmatrix} 1 & 1 & \delta \\ 1 & \delta & \delta \\ \delta & \delta & \delta \end{pmatrix}$
in	$\begin{pmatrix} \delta & 0 & \delta \\ \delta & \delta & 0 \\ \delta & \delta & \delta \end{pmatrix} \vee \begin{pmatrix} \delta & \delta & \delta \\ 0 & \delta & \delta \\ \delta & 0 & \delta \end{pmatrix}$

Kaavio 1. Ylimmän käsitetason topologisten suhteiden kuvaus 9-intersection matriisilla (δ kuvaa mitä tahansa arvoa: 0 tai 1).

Ylin käsitehierarkian taso on hyvin abstrakti eikä tarjoa geometrista yksityiskohtaista tietoa liittyen laajoihin reunusalueisiin ja useisiin komponentteihin. Keskimmäinen taso tarjoaa yksityiskohtaisemman kuvauksen perustuen 14 relaation. Taulukossa 2 on esimerkin vuoksi esitetty ylimmän tason relaatioista *touch* ja *in* polveutuvat keskimmäisen käsitetason relaatiot. Taulukko 3 esittää yhteenvedon kolmitasoisesta topologisten suhteiden käsitehierarkiasta.

nearlyInside	$\begin{pmatrix} \delta & 0 & \delta \\ 1 & \delta & 0 \\ \delta & 1 & \delta \end{pmatrix} \vee \begin{pmatrix} \delta & 0 & \delta \\ \delta & \delta & 0 \\ 1 & 1 & \delta \end{pmatrix}$
nearlyEqual	$\begin{pmatrix} \delta & 1 & \delta \\ 0 & \delta & 0 \\ \delta & 0 & \delta \end{pmatrix} \vee \begin{pmatrix} \delta & 0 & 0 \\ 0 & \delta & \delta \\ \delta & 0 & \delta \end{pmatrix} \vee \begin{pmatrix} \delta & 0 & \delta \\ \delta & \delta & 0 \\ \delta & 0 & \delta \end{pmatrix} \vee \begin{pmatrix} \delta & 0 & \delta \\ 0 & \delta & 0 \\ 0 & \delta & \delta \end{pmatrix}$

nearlyContains	$\begin{pmatrix} \delta & 1 & \delta \\ 0 & \delta & 1 \\ \delta & 0 & \delta \end{pmatrix} \vee \begin{pmatrix} \delta & \delta & 1 \\ 0 & \delta & 1 \\ \delta & 0 & \delta \end{pmatrix}$
nearlyMeet	$\begin{pmatrix} 0 & \delta & 1 \\ \delta & 1 & \delta \\ 1 & \delta & \delta \end{pmatrix}$
coveredByBoundary	$\begin{pmatrix} 0 & \delta & 0 \\ \delta & \delta & \delta \\ 1 & \delta & \delta \end{pmatrix}$
coversWithBoundary	$\begin{pmatrix} 0 & \delta & 1 \\ \delta & \delta & \delta \\ 0 & \delta & \delta \end{pmatrix}$
boundaryOverlap	$\begin{pmatrix} 0 & \delta & 0 \\ \delta & \delta & \delta \\ 0 & \delta & \delta \end{pmatrix}$

Kaavio 1. Keskimmäisen käsitetason topologisten suhteiden kuvaus 9-intersection matriisilla

Taso					Operaattoreiden lkm
Ylin taso (Top level)	disjoint	touch	overlap	in (and reverse)	4
Keskimmäinen taso (Intermediate level)	disjoint	nearlyMeet coveredByBoundary coversWithBoundary boundaryOverlap	nearlyOverlap interiorCoveredByInterior interiorCoversInterior partlyInside partlyContains crossCointainment	nearlyInside nearlyContains nearlyEqual	14
Alin taso (Bottom level, 9-Intersection)	1 relation matrix	16 relation matrices	26 relation matrices	13 relation matrices	56

Taulukko 3. Topologisten relaatioiden kolmitasoinen hierarkia

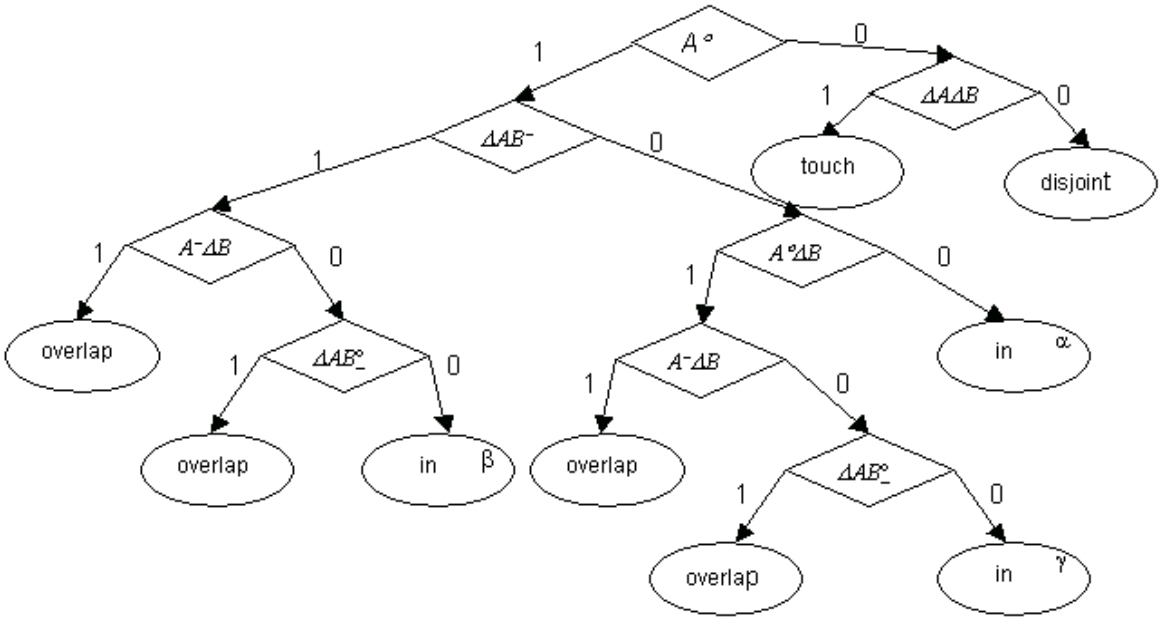
3.3.3 Päätöspuun käyttö predikaattien arvojen tutkinnassa

Päätöspuuta voidaan käyttää apuna selvitettäessä predikaattien voimassaoloa eri käsitehierarkiatasoilla [CFK00]. Puun solmuissa on 9-intersection matriisin sisältämien leikkausten testit. Näiden leikkausten arvoihin perustuen hakualuetta jaetaan niin, että lopulta puun lehtisolmu sisältää ainoastaan yhden relaation.

Päätöspuu on tarkoituksenmukaista rakentaa siten, että se minimoi laskennasta aiheutuvat kustannukset. Relaatiot tulisi siis pystyä löytämään mahdollisimman pienen testimäärän kautta. Päätöspuun rakentamisen pienimmällä mahdollisella testien määrällä on kuitenkin käytännössä

mahdotonta. Päätöspuualgoritmien päämääränä on pyrkiä minimoimaan puun koko, koska pienimmät puut mahdollistavat yleensä tarkemman luokittelun ja nopeamman laskennan.

Clementini, Di Felice ja Koperski [CFK00] käyttivät päätöspuun rakentamiseen ID3 algoritmia, joka pyrkii muodostamaan puun, jonka solmujen määrä on mahdollisimman pieni. Perusajatuksena on, että algoritmi tutkii kaikki leikkaukset ja jakaa puun perustuen aina leikkaukseen, jonka kautta saadaan eniten informaatiota. Kuvassa 2 on päätöspuu, joka on rakennettu oletuksella, että ylimmän tason relaatiot tunnetaan ja kaikki 56 alimman tason relaatiota ovat yhtä mahdollisia.



Kuva 3. Esimerkki päätöspuusta ylimmän tason relaatioille (A°B° tarkoittaa leikkausta A°∩B° jne.)

Päätöspuu voidaan rakentaa mille tahansa topologisen käsitehierarkian tasolle. Kun topologiset relaatiot on määritelty järjestyksessä, voidaan puun muodostaminen aloittaa ylimmältä tasolta ja edetä siinä alemmille tasoille. Esimerkiksi kuvassa esitetyn puun lehdistä kolme sisältää relaation *in*. Solmu α sisältää keskimmäisen tason relaatiot *nearlyInside ja nearlyEqual*, solmu β *nearlyContains* ja *nearlyEqual* ja solmu γ relaation *nearlyEqual*. Tunnettaessa ylimmän tason relaatiot keskimmäisen tason relaatioiden selvittämiseksi päätöspuu tarvitsee siis rakentaa vain solmuille α ja β.

3.3.4 Esimerkki käsitehierarkiaa hyödyntävästä louhinta-algoritmista

Spatiaalisia ja ei-spatiaalisia käsitehierarkioita voidaan käyttää apuna louhittaessa spatiaalisia assosiaatiosääntöjä spatiaalisesta tietokannasta, jossa objekteilla on laajat reunusalueet. Clementinin, Di Felicen ja Koperskin [CFK00] käyttämä Koperskin ja Hanin [KoH95] algoritmiin perustuva louhinta-algoritmi noudattaa seuraavia perusaskeleita:

1. Tallennetaan tehtävän kannalta oleelliset objektit *Task_relevant_DB*-tietokantaan.
2. Selvitetään MBR predikaatin arvot. MBR (Minimum Bounding Rectangles) kuvaa objektien välisiä relaatioita pienimpien mahdollisten objekteja ympäröivien suorakaiteiden tasolla. Mikäli objektien välillä on topologinen relaatio myös näiden objektien MBR:llä on oltava yhteisiä pisteitä. MBRiin perustuen voidaan siis suodattaa esiin predikaatit *in, overlap ja meet.*
3. Suodatetaan tulosjoukosta ne predikaatit, joihin liittyvien havaintojen määrä ylittää niille asetetun alarajan.
4. Jokaisella käsitehierarkian tasolla selvitetään predikaatit, joihin liittyvien havaintojen määrä on riittävän suuri ja louhitaan esiin niihin liittyvät assosisaatiosäännöt.

4. YHTEENVETO

Spatiaalisten assosiaatiosääntöjen louhinnan avulla selvitään spatiaalisessa tietokannassa usein esiintyviä spatiaalisten objektien keskinäisiä tai spatiaalisten ja ei-spatiaalisten objektien välisiä riippuvuuksia. Spatiaalinen tieto on luonteeltaan usein monitasoista ja siksi louhintamenetelmät pyrkivät etsimään assosiaatiosääntöjä useilta abstraktiotasoilta. Säännön esiintyessä ylemmällä tasolla on se yleensä löydettävissä myös muilta abstraktiotasoilta.

Louhittaessa spatiaalisia assosiaatiosääntöjä tehdään yleensä vaativimmat laskentaoperaatiot suuremmalla abstraktiotasolla ja sen jälkeen keskitetään laskenta objekteihin, jotka näin saadun karkean erottelun perusteella vaikuttavat mielenkiintoisimmilta. Tärkeässä asemassa spatiaalisia assosiaatiosääntöjä louhittaessa ovat myös käsitehierarkiat ja muu kohdealueeseen liittyvä

taustatieto, joka on hyödynnettävissä laskentaa tehtäessä. Koska spatiaaliseen tietoon liittyy usein epävarmuustekijöitä, tulisi louhintatekniikan pystyä käsittelemään luotettavasti myös epävarmaa tietoa. Ottamalla käyttöön laajan reunusalueen käsite voidaan spatiaaliseen tietoon usein liittyviä epävarmuustekijöitä suodattaa pois ja näin mahdollistaa tietoon liittyvät laskentaoperaatiot ilman että tietoon tarvitsee tehdä karkeita yksinkertaistuksia.

Lähteet:

[BKS93] Brinkhoff T., Kriegel H.-P., Seeger B.:
'Efficient Processing of Spatial Joins Using R-trees', Proceedings ACM SIGMOD International Conference on Management of Data, Washington, DC, 1993, 237-246.

[CFK00] Clementini, E. Di Felice, P. Koperski, K. Han, J.:
Mining multiple-level spatial association rules for objects with a broad boundary. Data & Knowledge Engineering, Volume 34, Issue 3, Sept 2000, 251-270

[KoH95] Koperski, K. Han, J.:
Discovery of Spatial Association Rules in Geographic Information Databases. Proc. 4th Int'l Symp. on Large Spatial Databases (SSD95)
http://www.cs.sfu.ca/people/GradStudents/koperski/personal/research/research.html

[HKS97] Han, J., Koperski, K., Stefanovic, N.:
GeoMiner: A System Prototype for Spatial Data Mining. Prec. ASM_SIGMOD Int. Conf. on Management of Data. SIGMOD Record 26(2), 1997, 553-556.
http://www.cs.sfu.ca/people/GradStudents/koperski/personal/research/research.html

[MaT97] Mannila, H. Toivonen, H: Levelwise search and borders of theories in knowledge discovery. Data Mining and Knowledge Discovery 1(3), 1997, 259-289.

[MEL01] Malerba, D. Esposito, F. Lisi, F. A:
Mining spatial association rules in census data. Proceedings of the Joint Conferences on New Techniques and Technologies for Statistcs and Exchange of Technology and Know-how (ETK-NTTS'01)
http://www.cs.sfu.ca/people/GradStudents/koperski/personal/research/research.html

[MEL02] Malerba, D. Esposito, F. Lisi, F. A.:
Mining spatial association rules in census data: A Relational Approach. In P. Brito and D. Malerba (Eds.), Notes of the ECML/PKDD 2002 Workshop on Mining Official Data, 80-93, Helsinki University Printing House: Helsinki. 2002.
http://lacam.di.uniba.it:8000/people/lisi.htm#Pubblicazioni

[PrS95] Preparata, F.P. Shamos, M. I. Computational Geometry: An Introduction, Springer-Verlag, 1995.