Anu Tanninen
Spatiaalisen tiedon louhinta -seminaari
Helsingin Yliopisto,
Tietojenkäsittelytieteen laitos
07.03.2003

Rikosten tutkiminen ja ehkäiseminen spatiaalisen tiedon louhinnan avulla

Rikoksista saatava tieto, joka talletetaan tietokantoihin, on määrältään valtava. Lisäksi käsittelemätön tieto on liian yksityiskohtaista, joten on hyödynnettävä tilastollisia analyysejä sen käsittelemiseksi [EsL01]. Näistä analyyseistä etsitään rikosten yleisyyteen vaikuttavia tekijöitä, joiden tunnistaminen on tärkeää ehkäistäessä uusia rikoksia. Hahmojen tunnistamiseen on käytetty muun muassa spatiaalisia tilastollisesti analyyttisiä metodeja (spatial statistical analytical method), jotka ovat laskennallisesti kalliita ja vaativat aiempaa tietoa sekä tietoa arvoalueesta (domain). Lisäksi metodit eivät tunnista suurissa spatiaalisissa tietokannoissa olevia tuntemattomia tai odottamattomia malleja kovinkaan helposti [EsL01].

Rikostapauskasojen (cluster) paikallistaminen on yksi ratkaisu, kun tutkitaan tietoköyhiä ympäristöjä. Tällaisissa ympäristöissä ei ole mahdollista pohtia muiden tasojen tietoja, jotka saattavat kuitenkin vaikuttaa kohdetasoon. Tietorikkaissa ympäristöissä puolestaan etsitään hahmoja kohdetasosta, ja tutkitaan sitten mahdollisia tekijöitä tai yhteyksiä, jotka perustuvat spatiaalisiin kasoihin. Klusterointi ja assosiaatiosääntöjen louhinta ovatkin tärkeitä louhittaessa spatiaalista tietoa.

Spatiaalinen klusterointi määritellään sarjana prosesseja, missä ryhmitellään joukko maantieteellisesti viitattua pistetietoa $P = \{p_1, p_2, ..., p_n\}$ jossakin tutkittavassa alueessa S pienempiin homogeenisiin aliryhmiin. Assosiaatiosääntöjen louhinnassa etsitään korrelaatioita suurista tietokannoista. Sääntö on ilmaus $X \Rightarrow Y$ *(c %)*, missä X on edeltävä tapahtuma ja Y on seuraus. X ja Y ovat vuorovaikutteisia tietokannassa ja niiden leikkaus on tyhjäjoukko. Säännön avulla ilmaistaan kuinka suuri osuus, *c %,* tiedosta, joka toteuttaa ehdon *X,* toteuttaa myös ehdon *Y* [EsL01]. Kullakin säännöllä on frekvenssi, joka ilmoittaa, kuinka usein hahmo esiintyy tiedossa sekä konfidenssi, joka ilmaisee *Y:*n esiintymistodennäköisyyden, kun X on annettu [MaT02]. Määriteltäessä minimifrekvenssi ja -konfidenssi, jotka sääntöjen on ylitettävä, saadaan esiin mielenkiintoiset ja merkittävät säännöt.

Monimutkaisten rikostapausten ymmärtämiseen käytetään rikosten kuumien pisteiden (hot spot) analyysejä [EsL01, GrM01, Lev02]. Näiden alueiden paikallistamiseen käytetään GIS-työkalua (Geografic Information Systems), joka yksinkertaistaa spatiaalisen tiedon keräämisen, käsittelyn ja analysoinnin tehden hahmojen tunnistamisen suoraviivaiseksi suuresta tietomäärästä [GAA99]. GIS:llä on kyky yhdistää spatiaalista tietoa muunlaiseen tietoon [GrM01]. Yleensä juuri tietynlaiset ympäristöt aiheuttavat enemmän rikoksia, mistä seuraa alueellista keskittymistä. Rikoksiin vaikuttavien tekijöiden etsimiseen käytetään spatiaalista tiedon louhintaa. Assosiaatiosääntöjen louhintaan on ehdotettu vertikaalista ja horisontaalista esitystä [EsL01]. Ensimmäinen tutkii useita maantieteellisiä tasoja, ja yrittää etsiä mielenkiintoisia yhteyksiä jonkin tietyn pisteen perusteella kustakin tasosta. Jälkimmäinen puolestaan asettaa kaikki tasot yhdelle kohdetasolle, josta se etsii yhteyksiä näiden eri tasojen leikkauspisteistä. Koska vertikaalisessa esitystavassa kukin taso jaetaan säännöllisiin soluihin, ovat löydetyt säännöt vahvasti riippuvaisia siitä, kuinka tasot on jaettu. Horisontaalisessa esityksessä ei tarvita lainkaan tietoa arvoalueesta. Lisäksi se on täysin itsenäinen ja sopiikin siksi paremmin suurten tietokantojen louhintaan kuin vertikaalinen esitys.

Rikoskartat ovat yksi tapa identifioida rikoksia visuaalisesti. Niistä saadaan informaatiota maantieteellisten alueiden, rikosten ja riskitekijöiden määrän välisistä suhteista. Carlos Carcach [Car99] tutki aseisiin liittyviä rikoksia Australiassa ja huomasi, että aseista johtuvat kuolemat ovat yleisempiä harvaan asutuilla alueilla. Miesten ja naisten riskit tulla ammutuksi tuottavat erilaisia alueellisia hahmoja. Esimerkiksi miesten kuolleisuus on yhteydessä maantieteelliseen eristyneisyyteen, kun taas naisten kuolleisuudella ei ole vastaavanlaista yhteyttä. Sukupuoleen perustuvat alueluokittelut ovatkin yksi mahdollisuus tutkia rikoksia, koska esimerkiksi australialaisten naisten aserikokset ovat yleisempiä eristyneillä alueilla, joissa harjoitetaan maanviljelyä. Lisäksi alueen ikärakenne vaikuttaa aserikosten määrään. Korkeisiin itsemurhalukuihin ovat vaikuttaneet muun muassa sosiaalinen pirstoutuminen, puute ja työttömyysluvut. Jotta rikoksia voitaisiin ennaltaehkäistä, käytetään spatiaalista tiedon louhintaa juuri tällaisten tapausten yhteydessä, jolloin voidaan löytää selviä syy-seuraus-yhteyksiä useammista eri rikoksista. Spatiaalinen tiedon louhinta onkin merkittävä prosessi hahmojen, yhteyksien, kasojen ja muiden ei-satunnaisten tapahtumien etsimiseen [Ans99]. CrimeStat [Lev02], joka on suunniteltu rikosten kartoitustutkimuskeskukselle, on spatiaalinen tilastollinen ohjelma, jolla voidaan analysoida näitä paikallistettuja rikostapauksia.

Rikosten ennaltaehkäisemiseksi on tutkittava syy-seuraus-yhteyksiä, jotta tiedettäisiin, mitkä tapahtumat johtavat yleisimmin erilaisiin rikoksiin. Esimerkiksi FBI käyttää spatiaalista tiedon louhintaa rikosten selvittämiseen [EsL01]. Louhinnan avulla voidaankin löytää mahdollisia tekijöitä, kuten asukastiheys ja maantieteellinen eristyneisyys, erilaisiin rikoksiin.

Lähteet

[Ans99]: Luc Anselin, The Future of Spatial Analysis in the Social Sciences. Geographic Information Sciences 5, 67-76

[Car99]: Carlos Carcach, Spatial Analysis of Crime Data: Firearms Related Homicide in Australia. Paper presented at the 3 rd National Outlook Symposium on Crime in Australia, Mapping the Boundaries of Australia's Criminal Justice System convened by the Australian Institute of Criminology and held in Canberra, 22-23 March 1999

[EsL01]: Vladimir Estill-Castro and Ickjai Lee, Data Mining Techniques for Autonomous Exploration of Large Volumes of Geo-referenced Crime Data. 6th International Conference on Geocomputation, 24-26, September, 2001, Brisbane, Australia

[GAA99]: Michael F. Goodchild, Luc Anselin, Richard P. Appelbaum, Barbara Herr Harthorn, Toward Spatially Integrated Social Science.International Regional Science Review 23, 139-159.

[GrM01]: Tony H. Grubesic, Alan T. Murray, Detecting Hot Spots Using Cluster Analysis and GIS. Proceedings from the Fifth Annual International Crime Mapping Research Conference. December 1, 2001, Dallas, TX.

[Lev02]: Ned Levine, CrimeStat A Spatial Statistics Program for the Analysis of Crime Incident Locations (Version 2.0) [Computerfile]. Houston, TX: Ned Levine & Associates/Washington, DC: U.S. Dept. of Justice, National Institute of Justice [producers], 2002. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2002.

[MaT02]: Heikki Mannila, Hannu Toivonen, Knowledge Discovery in Databases: Search for Frequent Patterns. Kurssin Tietämyksen muodostaminen, syksy 2002 kurssimateriaali