LUCRĂRILE SIMPOZIONULUI "SISTEME INFORMAȚIONALE GEOGRAFICE", NR. 10
Anal. Șt. Univ. „Al.I.Cuza" Iași, Tom L, s.IIc. Geografie, 2004

DESCOPERIREA DE CUNOȘTINȚE DIN BAZE DE DATE SPAȚIALE

Mirela Danubianu*

În perioada ultimilor ani am asistat la un proces de stocare a unor cantități uriașe de date, în baze de date, depozite de date, sisteme informaționale geografice sau sub alte forme. Este evident faptul că, în prezent, aceste cantități de date cresc cu o viteză uluitoare.

Numeroase organizații și-au construit depozite de date capabile să stocheze sute de teraocteți de date cu privire la explorarea resurselor naturale; bazele de date din domeniul astronomiei au dimensiuni de ordinul teraocteților; se estimează că sistemele de observare a Pământului pot transmite până la 50 gigaocteți pe oră. Aceste aspecte au constituit o provocare pentru metodele tradiționale de analiză a datelor capabile să extragă informații și cunoștințe.

In contextul dat, explorarea datelor (DM - data mining) și descoperirea cunoștințelor din date (KDD - knowledge discovery in databases) s-au impus ca domenii de cercetare de mare interes. In esență, *acestea implică descoperirea de cunoștințe interesante, anterior necunoscute și implicite din baze mari de date* [FPSU96], și se află la confluența mai multor domenii de cercetare, incluzând învățarea automată, sisteme de baze de date, statistică, recunoașterea formelor și teoria informației.

In afara numeroaselor studii de descoperire a cunoștințelor din bazele de date relaționale, sau din bazele de date de tranzacții, explorarea datelor spațiale (spatial data mining), care se referă la *extragerea cunoștințelor implicite, a relațiilor spațiale sau a altor tipare care nu sunt explicit memorate în bazele de date spațiale,* a făcut obiectul a numeroase studii de cercetare [EKX95][KN96][KH95][ZRL96].

Datele spațiale au o serie de caracteristici care le disting de bazele de date relaționale. In principal este vorba despre faptul că acestea conțin informații topologice sau de distanță, folosesc structuri de indexare spațiale multidimensionale, sunt accesate prin metode de acces specifice și solicită adesea calcule geometrice și tehnici spațiale de reprezentare a cunoștințelor. Ca urmare, explorarea datelor spațiale necesită integrarea tehnologiilor de explorare a datelor cu cele specifice bazelor de date spațiale. Una din problemele delicate ale explorării datelor spațiale este aceea a eficienței tehnicilor folosite, datorită faptului că se lucrează cu volume importante de date de tipuri complexe și cu metode de acces spațiale .

Explorarea datelor spațiale poate fi utilizată pentru vizualizarea bazelor de date spațiale, pentru înțelegerea acestui tip de date, pentru a descoperi legăturile spațiale și relația dintre datele spațiale și cele non-spațiale, pentru reorganizarea bazelor de date spațiale, optimizarea interogărilor spațiale, etc. Rezultatele acestor cercetări își găsesc aplicația în sistemele informaționale geografice, în explorarea bazelor de date de tipul imaginilor, în navigație sau în orice alte domenii care lucrează cu date spațiale.

Această lucrare se dorește a fi o scurtă trecere în revista a metodelor de explorare a datelor spațiale, a tehnicilor utilizate, cu plusurile și minusurile lor, cu modurile lor de aplicare și cu provocările cărora trebuie să le facă față.

* Universitatea "Ștefan cel Mare" Suceava, Facultatea de Inginerie Electrică

Utilizarea tehnicilor de explorare a bazelor de date relaţionale

Una din primele încercări de analiză a datelor spaţiale a fost realizată în 1993 de către Major et al. [MM93] care au utilizat un instrument comercial pentru a explora o bază de date referitoare la o furtună tropicală, în scopul de a prezice dacă aceasta va atinge teritoriile Statelor Unite ale Americii. Datele prin care se descria uraganul au fost descompuse în observaţii în diferite puncte, iar aceste observaţii au fost memorate într-o bază de date relaţională. Au fost folosite atribute de tipul: poziţia uraganului, viteza, direcţia de deplasare, unghiul faţă de coastă, etc. Deoarece descrierea uraganului în diferite puncte a fost memorată prin intermediul mai multor tuple, anumite date au fost interdependente. Aceste interdependenţe au provocat o serie de neajunsuri deoarece algoritmul folosit presupunea independenţa datelor. Ca suport al selecţiei pentru cele mai bune reguli a fost utilizat un sistem GIS.

Concluziile acestui studiu au fost acelea că este necesară extinderea tehnicilor tradiţionale de explorare a datelor către explorarea datelor spaţiale pentru o analiză corespunzătoare a obiectelor şi a fenomenelor spaţiale complexe.

Procesul descoperirii de cunoştinţe din bazele de date spaţiale

Experienţa ultimilor câţiva ani a arătat faptul că descoperirea cunoştinţelor din baze de date uriaşe implică mai mult decât simpla aplicare a algoritmilor sofisticaţi de explorare a datelor pe un set predefinit de date.

Una din problemele majore în cercetările legate de descoperirea cunoştinţelor din date este aceea a înţelegerii KDD ca un ***"proces netrivial de identificare a tiparelor valide, cu caracter de noutate, potenţial folositoare şi în ultimă instanţă posibil de înţeles, din date."*** [FPS96]

Din acest punct de vedere, noţiunea de tipar este înţeleasă într-un sens foarte general. Un tipar este tot ceea ce un algoritm de explorare a datelor poate extrage sau poate genera din date, cum ar fi un model care clasifică pe baza unui arbore de decizie sau a unei reţele neuronale, o grupare de date sau o mulţime de reguli de asociere.

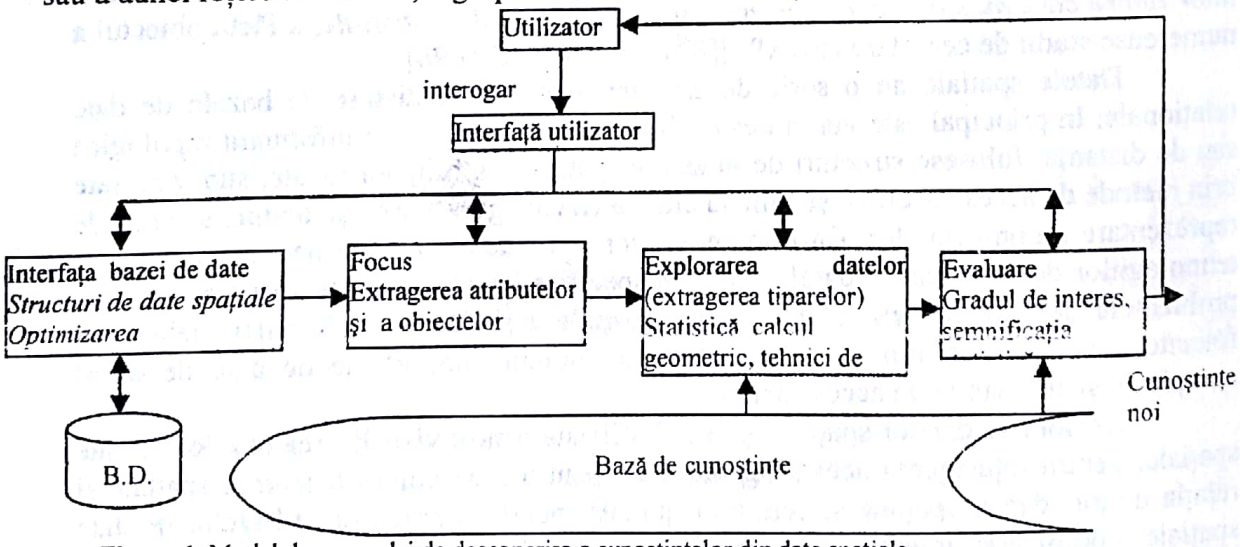

Figura 1 Modelul procesului de descoperire a cunoştinţelor din date spaţiale

Procesul de descoperire a cunoştinţelor din date spaţiale poate fi modelat prin arhitectura[†] prezentată în figura 1.

[†] Adaptare după Matheus, C.J. , Chan, P.K., şi Piatetsky-Shapiro, G. Systems for Knowledge Discovery in Databases. IEEE Transaction Knowledge and Data Engineering 5,5 (oct. 1993), 903-913

Experienţa acumulată a dus la concluzia că procesul descoperirii cunoştinţelor nu se reduce la o „simplă apăsare a unui buton", ci din contră este complex, iterativ şi puternic interactiv. In fiecare din etapele sale se face simţită prezenţa omului prin utilizatorul (sau analistul) care decide în ce va consta următoarea fază, dacă va fi reluată faza curentă sau chiar dacă se va face o întoarcere un pas înapoi la o fază anterioară.

Pe scurt, descoperirea cunoştinţelor se realizează astfel: cunoştinţele de fond, cum ar fi ierarhia de concepte spaţiele şi non-spaţiale sau informaţiile legate de baza de date sunt memorate în *baza de cunoştinţe*. Datele sunt extrase de pe mediul de stocare prin intermediul *interfeţei bazei de date* care permite, printre altele, optimizarea interogărilor. Pot fi folosiţi indecşi pentru datele spaţiale, în scopul procesării eficiente. *Componenta de focalizare* decide care sunt acele părţi ale datelor utile pentru recunoaşterea tiparelor Spre exemplu, se poate decide că numai anumite atribute sunt relevante pentru sarcina de descoperire a cunoştinţelor sau se pot extrage obiecte a căror utilizare garantează rezultatele cele mai bune. Regulile şi tiparele sunt descoperite în etapa aplicării tehnicilor de *explorare a datelor* realizată de modulul cu acelaşi nume. Acesta poate utiliza procedee statistice, metode de învăţare automată şi tehnici de clasificare şi grupare în conjuncţie cu algoritmi de calcul geometric, pentru descoperirea de reguli sau legături. Importanţa şi semnificaţia tiparelor descoperite este stabilită cu ajutorul modulului de *evaluare* care are posibilitatea de a elimina cunoştinţele deja existente sau redundante. Cunoştinţele descoperite sunt oferite în final utilizatorului pentru a fi verificate.

Metode şi tehnici de explorare a datelor folosite

Cunoştinţele descoperite din bazele de date spaţiale pot îmbrăca diverse forme, şi anume: reguli caracteristice pentru descrierea datelor spaţiale, reguli discriminante pentru a diferenţia o clasă de date spaţiale de alte clase, reguli de asociere care asociază una sau mai multe caracteristici unui set distinct de caracteristici, reguli de deviaţie şi de evoluţie pentru descrierea schimbărilor în timp sau reguli care descriu structuri deosebite. Toate acestea pot fi prezentate sub diferite forme şi pot fi utilizate pentru a descrie obiectele spaţiale.

Deoarece explorarea datelor impune numeroase aspecte provocatoare pentru cercetare, aplicarea directă a metodelor şi tehnicilor dezvoltate de domenii înrudite precum învăţarea automată, statistica sau bazele de date nu pot rezolva problemele apărute. In acest scop este necesară efectuarea de studii dedicate pentru a găsi noi metode sau tehnici integrate pentru o explorare eficientă. Ca urmare a acestui fapt, explorarea datelor (data mining) s-a constituit ea însăşi într-un domeniu aparte de cercetare.

In calitate de componentă a procesului de descoperire a cunoştinţelor din date, explorarea datelor implică aplicarea iterativă a unor metode specifice, în particular a unor algoritmi specifici.

Se pot distinge două tipuri de obiective, ale acestei etape, şi aume: *verificarea*, caz în care sistemul este folosit pentru a verifica ipotezele utilizatorului şi *descoperirea* atunci când sistemul găseşte, în mod autonom noi tipare. Mai departe, descoperirea poate fi divizată în : *predicţie* dacă. sistemul găseşte tipare în scopul predicţiei comportamentului viitor pentru anumite entităţi sau *descriere* atunci când sistemul găseşte tipare în scopul prezentării acestora unui utilizator, într-o formă inteligibilă.

Obiectivele predicţiei respectiv ale descrierii sunt realizate prin intermediul următoarelor metode primare de explorare a datelor:

- **clasificarea** : găsirea unei funcţii care include un articol de date într-una din mai multe clase predefinite.
- **regresia**: este utilizată la prezicerea unei valori a unei variabile continue bazată pe valorile altor variabile, presupunând un model de dependenţă liniar sau neliniar. Regresia logică este utilizată pentru precizarea valori unei variabile binare. Este un instrument de clasificare care este utilizat la precizarea valorii unei variabile, cum ar fi de exemplu, "dacă un individ este cumpărător sau nu", deasemenea este utilizată la precizarea variabilelor continue, cum ar fi: "probabilitatea că un individ va face cumpărături".
- **gruparea** (clusering-ul): identifică o mulţime finită de categorii sau cluster-e pentru a descrie datele.Strâns legată de acesta este este metoda estimării densităţii de probabilitate care constă în tehnici de estimare din date a funcţiei de probabilitate asociată pentru toate variabilele/câmpurile unei baze de date.
- **rezumarea:** găseşte o descriere compactă pentru o submulţime de date.
- **modelarea dependenţelor:** găseşte un model care descrie dependenţele semnificative dintre variabile.
- **detectarea schimbărilor şi a deviaţiei :** descoperă cele mai semnificative schimbări produse în date în intervalul dintre două măsurări consecutive.

Se poate face o clasificare după fundamentul căii de explorare a datelor în explorare pe baza de generalizare, explorare bazată pe tipare, explorare bazată pe teoriile matematice şi statistice, acces integrat etc

Explorarea pe bază de generalizare

Datele şi obiectele din bazele de date conţin adesea informaţii detaliate la nivelul conceptelor primitive. In acest caz este de dorit să facă o rezumare a mulţimilor de date pentru prezentarea acestora prin concpte de nivel superior. Se pot, spre exemplu, rezuma datele detaliate despre temperaturi şi precipitaţii dintr-o regiune în scopul prezentării tiparului general al vremii. Aceasta implică o explorare a datelor bazată pe generalizare, caz în care în prima fază are loc o abstractizare a unui volum considerabil de date relevante de la un concept de nivel scăzut la unul de nivel relativ ridicat, după care se realizează extragerea cunoştinţelor din datele generalizate. In această situaţie se impune existenţa cunoştinţelor fundamentale în forma unei ierarhii de concepte, care este fie furnizată explicit de experţii în domeniu, fie poate fi generată automat prin analiza datelor.

In cazul bazelor de date spaţiale pot fi definite două categorii de ierarhii de concepte: *non-spaţiale şi spaţiale.* Odată cu impunerea ierarhiilor de concepte, informaţia devine din ce în ce mai generală, dar este încă consistentă la nivelul conceptelor inferioare. In cazul datelor spaţiale, un exemplu de ierarhie în procesul de generalizare ar fi următorul: regiunile reprezentând unităţi teritorial-administrative pot fi unite în judeţe, iar acestea la rândul lor pot fi unite în state ş.a.m.d.

Inducţia orientată spre atribute este o tehnică eficientă de generalizare a datelor [FPSU96]. Se consideră mai întâi o interogare de explorare a datelor exprimată într-un limbaj similar SQL,(de exemplu DMSQL) care colectează setul de date relevante într-o bază de date. Apoi generalizarea este realizată prin creşterea ierarhiei de generalizare şi rezumarea legăturilor dintre datele spaţiale şi cele non-spaţiale la concepte de nivel superior. In cazul datelor non-spaţiale aceasta se realizează prin: ridicarea ierarhiei de concepte când valorile atributelor dintr-un tuplu sunt înlocuite prin valorile generalizate, prin eliminarea atributelor atunci când continuarea generalizării este imposibilă şi există prea multe valori distincte pentru un atribut şi prin cuplarea

tuplelor identice. Inducţia continuă până când toate atributele sunt generalizate până la nivelul dorit. Pe durata procesului de unire a tuplelor identice numărul tuplelor cuplate este memorat ca un contor. Suplimentar pot fi memorate valorile agregate ale unor atribute cantitative pentru a face posibilă prezentarea cantitativă a cunoştinţelor obţinute.

Datele generalizate pot fi exprimate sub forma unor relaţii generalizate sau cuburi de date asupra cărora pot fi aplicate alte operaţii în scopul transformării lor în diverse forme de cunoştinţe. Spre exemplu oparaţiile de "drill-down" respectiv de "roll-up" permit vizualizarea datelor la diferite nivele de abstractizare; relaţiile generalizate pot fi transpuse în tabele rezumate, hărţi sau diagrame în scopul vizualizării şi prezentării, pot fi extrase reguli caracteristice sau reguli discriminante, etc.

Inducţia orientată spre atribute a fost extinsă şi asupra datelor spaţiale. In [LHC93] sunt prezentaţi doi algoritmi de generalizare : în cazul dominanţei datelor spaţiale respectiv în cazul dominanţei datelor non-spaţiale.

Algoritmul de generalizare în cazul dominanţei datelor spaţiale permite descrierea regiunilor din spaţiu cu ajutorul predicatelor de nivel înalt. In primul rând sunt unite regiunile în concordanţă cu ierarhiile spaţiale ceea ce conduce la o hartă cu un număr redus de zone. Apoi, se realizează o descriere non-spaţială a fiecărei zone cu utilizând tehnica inducţiei orientate pe atribute. Răspunsul unei interogări este descrierea tuturor regiunilor cu ajutorul disjuncţiei câtorva predicate care caracterizează fiecare din regiunile generalizate.

Algoritmul de generalizare pentru date dominante non-spaţiale crează hărţi care constau dintr-un număr mic de regiuni care se află pe acelaşi nivel ierarhic al descrierii nonspaţiale. Acest algoritm începe cu inducţia orientată către atribute nonspaţiale şi generalizarea acestora către nivele conceptuale mai înalte. Apoi zonele învecinate având aceleaşi valori pentru atributele generalizate sunt unite. Se poate utiliza aproximarea pentru a ignora regiunile reduse cu descrieri non-spaţiale diferite. In figura următoare este ilustrat un exemplu al aplicării algoritmilor menţionaţi. Este prezentată interogarea formulată în algoritmul de generalizare cu dominanţă spaţială precum şi rezultatul obţinut.

```
Extract characteristic rule
 from harta_temperaturi
   where provincia ="XX" and
   perioada="vara" and anul=1999
      in relevance to regiune and
temperatură
```

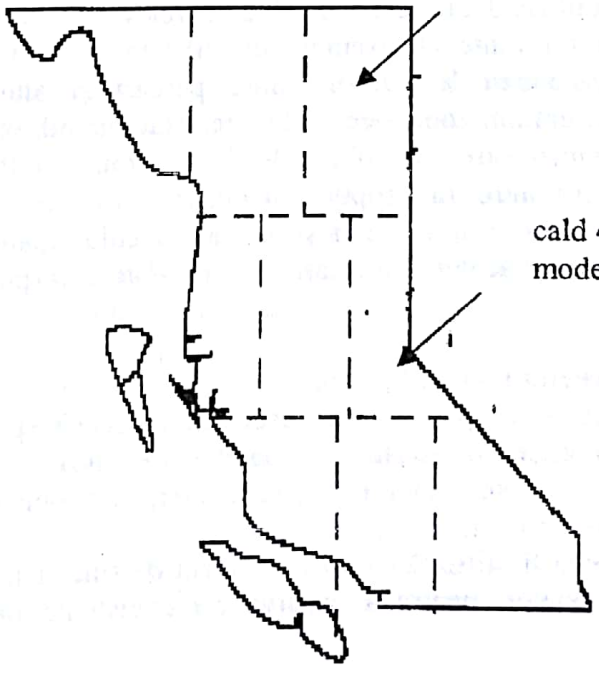

Figura 2. Interogarea şi rezultatul obţinut în urma aplicării algoritmului de generalizare cu dominanţă spaţială

Explorarea asocierilor spaţiale

Metoda amintită anterior permite găsirea regulilor caracteristice care descriu obiectele spaţiale din punct de vedere al atributelor lor non-spaţiale. Sunt cazuri în care, însă, se doreşte descoperirea regulilor de asociere spaţiale, care sunt acele reguli care asociază unul sau mai multe obiecte spaţiale cu alte obiecte spaţiale.

Conceptul de reguli de asociere a fost introdus în 1993 [AIS93] pentru explorarea bazelor de datede tranzacţii de dimensiuni mari, iar în 1995 [KH95]a fost extins la bazele de date spaţiale.

O regulă de asociere spaţială se notează :

$$X \rightarrow Y\ (c\%)$$

unde: X şi Y sunt mulţimi de predicate spaţiale sau non-spaţiale iar
c% este gradul de încredere (confidenţa) regulii

Un exemplu de astfel de regulă este următoarea:

este (x, şcoală) ∧ aproape_de (x, centru_sportiv) → aproape_de (x,parc) (80%)

Această regulă afirmă că 80% din şcolile care sunt aproape de un centru sportiv sunt de asemenea aproape de un parc. Există diferite tipuri de predicate spaţiale care pot genera reguli de asociere spaţiale, printre care: relaţii topologice precum *intersecţia, suprapunerea, disjuncţia* etc., sau informaţii de distanţă, precum *aproape_de, îndepărtat_de*, etc.

In bazele de date mari pot exista numeroase asociaţii între obiecte dar cele mai multe din acestea, fie sunt aplicabile doar unui număr redus de obiecte, fie au un grad de încredere redus.

Spre exemplu, un utilizator poate să nu fie interesat de o relaţie care asociază 5% din case cu o anumită şcoală, dar poate fi interesat de o regulă care se aplică cel puţin la 50% din case. Există două praguri, *suportul minim şi gradul minim de încredere* care sunt utilizate pentru controlul filtrării asocierilor care descriu un procent mic de obiecte sau regulile cu confidenţă redusă. Aceste praguri pot fi diferite la fiecare nivel al descrierilor non-spaţiale ale obiectelor, deoarece utilizarea aceluiaşi prag poate să nu găsească asociaţiile interesante la nivel ierarhic redus pentru că numărul obiectelor care satisfac acelaşi predicat poate fi destul de mic.

Procesul de explorare este iniţiat printr-o interogare care descrie o clasă de obiecte S utilizând alte clase de obiecte relevante pentru sarcina propusă şi o mulţime de legături relevante. De exemplu, un utilizator poate dori să descrie parcurile unui oraş prin reprezentarea legăturilor dintre parcuri şi alte obiecte precum: căi ferate, restaurante, grădini zoologice, străzi etc. Mai curând, utilizatorul poate stabili că este interesat numai care sunt obiectele dintr-o zonă cu lăţimea de un kilometru de la marginile parcului. In scopul minimizării costului calculelor spaţiale, algoritmul utilizează diferite aproximări şi aplică calculul spaţial mai detaliat, dar mult mai costisitor, numai acelor tipare care au un suport corespunzător la nivelul de aproximaţie considerat.

Gruparea datelor (Data Clustering)

Gruparea spaţială, care asociază obiectele spaţiale similare în clase, este o tehnică importantă de explorare a datelor, care poate fi utilizată în scopul identificării zonelor de exploatare similară a resurselor, în fuzionarea regiunilor cu caracteristici climatice similare, etc.

Poate fi utilizată ca un instrument de sine stătător pentru a înţelege mai bine distribuţia datelor, pentru a observa caracteristicile fiecărui grup şi pentru a uşura

focalizarea asupra unei anumite mulţimi de grupuri pentru analize ulterioare, sau, poate fi un pas de pregătire pentru algoritmii de clasificare sau de caracterizare care operează asupra grupurilor detectate.

Urmare a faptului că analiza grupurilor a fost un domeniu de cercetare activ în explorarea datelor, în ultimii ani au fost dezvoltate metode eficiente de grupare.Acestea pot fi clasificate în: metode de partiţionare [KR90][NH94][BFR98], metode ierarhice [KR90] [ZRL96] [GRS98] [KHK99], metode bazate pe densitate [EKSX96] [ABKS99] [HK98], metode bazate pe grilă [WYM97] [SCZ98] [AGGR98] şi metode bazate pe model [SD90] [KOH82].

Utilizarea aproximărilor şi a agregărilor

Metodele de grupare pot oferi răspunsuri la întrebări de genul "unde sunt poziţionate grupurile în bazele de date spaţiale". Există însă şi un alt aspect al problemei, şi anume "de ce se găsesc grupurile acolo". Acesta poate fi reformulat astfel: care sunt caracteristicile grupurilor, prin prisma proprietăţilor sau a obiectelor pe care le conţin acestea. Problema care se pune este legată de modul de măsurare a proximităţii agregării deoarece o afirmaţie de genul "90% din casele din grup posibil să aibă caracteristica X" este mai interesantă şi conţine mai multă informaţie decât afirmaţia "o casă are cu siguranţă caracteristica X". Proximitatea agregării este măsura apropierii setului de obiecte din grup de o anumită caracteristică ca opusă distanţei dintre graniţele grupului şi limitele unei caracteristici.

În [KN96] este prezentată o metodă care permite găsirea rapidă a unei caracteristici apropiată grupului.

Explorarea bazelor de date raster

Explorarea unor baze de date raster, precum cele de tipul imaginilor, este un caz particular al explorării datelor spaţiale. Printre studiile în domeniu se numără POSS II (Second Palomar Observatory Sky Survey) care utilizează arborii de decizie pentru clasificarea galaxiilor, sau a altor obiecte cosmice pornind de la cca 3 teraocteţi de imagini cosmice şi studiul Magellan care analizează aproximativ 30000 de imagini radar de rezoluţie înaltă ale suprafeţei planetei Venus, în scopul identificării posibililor vulcani.

Provocări pentru viitor

Este inutil, ca în anii de tranziţie către societatea informaţională se se demonstreze necesitatea unui domeniu de cercetare precum cel al extragerii de cunoştinţe din volume mari de date geografice.

Ceea ce trebuie însă menţionat este faptul că există încă aspecte ale cercetării care necesită extinderea studiilor. Câteva din aceste aspecte se referă la: găsirea unor metode adiţionale precum clasificarea, explorarea bazată pe tipare sau pe similarităţi special dedicate datelor geografice; analiza posibilităţilor de explorare interactivă a datelor folosind reacţiile vizuale, şi analiza posibilităţilor de proiectare a unui limbaj dedicat explorării datelor spaţiale; explorarea datelor provenite din surse distribuite (Internet/Intranet) şi memorate în diferite formate. O certă provocare este combinarea metodelor de explorare a datelor spaţiale cu bazele de date spaţiale avansate (baze de date spaţiale orientate obiect sau baze de date spaţio-temporale)şi cu tehnologia sistemelor expert pentru a crea aşa-numitele sisteme GIS inteligente.

Concluzii

Am făcut în această lucrare o trecere în revistă a metodelor utilizate în explorarea datelor, văzută ca o etapă a procesului de descoperire a cunoştinţelor din bazele de date spaţiale. Este o temă actuală şi în acelaşi timp atractivă care are ca scop să ofere sistemelor informaţionale geografice puterea descoperirii cunoştinţelor şi să adapteze noile aplicaţii GIS cerinţelor anilor ce vor veni.

REFERINŢE

[ABKS99]Ankerst, M., Breunig, M., Kriegel, H.P., Sanders, J., *OPTICS: Ordering points to identify the clustering structure*. In Proc. 1998 ACM- SIGMOD Int. Conf. Management of Data (SIG-MOD'99) Philadelphia, 1999, 49-60

[AIS93] Agrawal, R., Imielinski, T. and Swami, A. (1993) *Mining association rules between sets of items in large databases*. In Proceedings of the ACM SIGMOD International Conference on Management of Data (ACM SIGMOD 1993)9

[AGGR98]Agrawal, R., Gehrke, J., Gunopulos, D., Raghavan, P., *Automatic subspace clustering of high dimensional data for data mining applications*. In Proc 1998 ACM-SIGMOD Int. Conf. Management of data (SIGMOD'98), Seattle, 1998, 94-105

[BFR98]Bradley, R., Fayyad, U., Reina, C., *Scaling clustering algorithm for large databases* In Proc. 1998 Int. Conf. Knowledge Discovery and Data Mining (KDD'98) New York, 1998, 9-15

[EKX95] Ester, M., Kriegel, HP., Xu, X., *Knowledge Discovery in Large Spatial Databases : Focusing Techniques for Efficient Class Identification*. In Advances in Spatial Databases, SSD'95, Springer-Verlag , Berlin, 1995, 67-82

[EKSX96] Easter, M., Kriegel, H.P., Sander, J.,Xu, X.*A density-based algorithm for discovering clusters in large spatial databases* In Proc. 1996 Int. Conf. Knowledge Discovery and Data Mining (KDD'96)Portland, 1996, 226-231

[FPS96] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996)*The KDD process for extracting useful knowledge from volumes of data*. Communications of the ACM, 39(11):27-34

[FPSU96]Fayyad, U., Piatetsky-Shapiro, G., Smith, P., Uthurusamy, R., *Advances in Knowledge Discovery and Data Mining* AAAI/MIT Press, Menlo Park, CA, 1996.

[GRS98] Guha, S., Rastogi, R., Shim, K., *Cune: An efficient clustering algorithm for large databases* In Proc 1998 ACM-SIGMOD Int. Conf> Management of Data (SIGMOD'98),Seattle, 1998, 73-84

[HK98] Hinneburg, A., Keim, A., *An efficient approach to clustering in large multimedia databases with noise*. In Proc 1998 Int. Conf. Knowledge Discovery and Data Mining (KDD'98), New York, 1998, 58-65

[KH95]Koperski, K., Han, J., *Discovery of Spatial Association Rules in Geographic Information Databases*, SDD'95 Springer-Verlag, Berlin, 1995, 47-66

[KHK99]Karypis, G., Han,J., Kumar, V., *CHAMELEON: A hierarhical clustering algorithm using dynamic modeling*. COMPUTER, 32:68-75, 1999

[KN96] Knorr, E., Ng, R. T.,*Finding aggregate Proximity Relationship and Commonalities in Spatial Data Mining*. IEEE Transaction Knowledge and Data Engineering 1996, 884-897

[KOH82]Kohonen, T., *Self organized.formation of topologically correct feature maps*. Biological Cybernetics, 1982, 43:59-69

[KR90] Kaufman, L., Rousseeuw, P.J., *Finding Groups in Data: an introduction to Cluster Analysis*, John Wiley & soons, 1990

[LHC93]Lu, W., Han, J., Choi, B.C., *Discovery of General Knowledge in Large Spatial Databases* In Proc. Of Far East Workshop on Geographic Information Systems, Singapore, 1993, 275-289

[MM93] Major, J., Mangano, J., *Selecting among Rules Induced from a Hurricane Database*. In Proc. 1993 Knowledge Discovery in Databases Workshop AAAI Press, 1993

[NH94]Ng, R., Han, J., *Efficient and effective clustering method for spatial data mining* In Proc. 1994 Int. Conf. Very Large Data Bases (VLDB94), Santiago Chile, 1994, 144-155

[SCZ98]Sheikholeslami, G., Chatterjee, S., Yhang, A., *WaveCluster:A multi-resolution clustering approach for very large spatial databases*, In Proc. 1998 Int Conf. Very Large Data Bases (VLDB98), New-York, 1998, 428-439

[SD90] Shavlik, J.W., Dietterich, T.G., *Reading in Machine Learning*, Morgan Kaufmann, 1990

[ZRL96] Yhang, T., Ramakrishnan, R., Livny, M., *BIRCH: an efficient data clustering method for very large databases*. In Proc 1996 ACM-SIGMOD Int. Conf. Management of Data, Montreal, 1996, 103-114

[WYM97] Wang, W. Yang, J., Munty, R., *STING: A statistical information grid approach to spatial data mining*. In Proc 1997 Int. Conf. Very Large Data Bases (VLDB'97) Athenes, 1997, 186-195