ASPECTE CHEIE ÎN GEOSPATIAL DATA MINING

dr. Angela Ioniţă

Institutul Naţional pentru Cercetare-Dezvoltare în Informatică – ICI, Bucureşti

Rezumat: Organizat în patru capitole, acest articol este un rezumat al Raportului Tehnic 630/2, elaborat în mai 2002, în cadrul Proiectului INTAS nr. 397: DATA MINING TECHNOLOGIES AND IMAGE PROCESSING: THEORY AND APPLICATIONS. În primul capitol, sunt prezentate aspecte legate de definiţia termenilor data mining şi *knowledge discovery* pentru ca, în capitolul al doilea, să se urmărească prezentarea caracteristicilor datelor geospaţiale, care trebuie luate în considerare în contextul *data mining*, în general, şi în contextul obiectivelor acestui proiect, în special. Următorul capitol este dedicat sistemelor software relevante pentru *geospatial data mining* (GeoMiner, ADaM, SPIN!). Ultimul capitol prezintă câteva concluzii privind impactul asupra ştiinţelor legate de informaţia geografică şi aspectele critice al cercetărilor în domeniu. Menţionăm, de asemenea, că acest articol, ca şi întregul Raport Tehnic 630/2, este o sinteză bazată pe articolele, cărţile şi prezentările menţionate în capitolul dedicat referinţelor bibliografice, pe care le-am considerat relevante pentru demersul din cadrul activităţilor Proiectului INTAS nr. 397.
Facem de la început precizarea că termenul "data mining" ar putea fi tradus ca "*mineritul datelor*". Deoarece până la data publicării acestui articol nu există consens asupra termenului în limba română, pentru a evita confuziile, îl vom utiliza pe cel în limba engleză. La fel vor fi utilizaţi şi ceilalţi termeni asupra cărora se mai lucrează în ceea ce priveşte completarea definiţiilor (datorită dinamicii cercetării) şi atribuirii unui termen românesc adecvat.

Cuvinte cheie: geographical information, geographical knowledge, Geographical Information Sistems (GIS), data mining/knowledge discovery.

Context

Acest articol este rezumatul Raportului tehnic 630/2, executat în cadrul activităţilor specifice Proiectului **INTAS no. 39**: **DATA MINING TECHNOLOGIES AND IMAGE PROCESSING: THEORY AND APPLICATIONS.**

Început în 1 noiembrie 2001, cu o durată de trei ani, obiectivele acestui proiect sunt:

- elaborarea metodelor pentru găsirea de şabloane, descrieri de date de volum mare şi construirea de modele predictive;
- elaborarea şi investigarea de metode de distilare şablon, bazate pe abordarea colectivă în clasificarea nesupervizată;
- dezvoltarea şi investigarea de noi metode de căutare şi analiză a dependenţelor neliniare complexe;
- elaborarea şi investigarea noilor metode pentru soluţionarea activităţilor dinamice în cadrul data mining:
- efectuarea de prognoze agricole, bazate pe studierea factorilor meteorologici şi pe geomonitorizarea regiunilor;
- modelarea curgerii râurilor folosind reţele neuronale;
- elaborarea de noi metode de discriminarea structurilor spaţiale locale, în cadrul tipurilor de date GIS;
- elaborarea de noi instrumente pentru analiza texturii bazate pe caracteristici universale şi morfologie matematică;
- elaborarea de modele de calcul adecvate pentru implementarea reţelelor neuronale ierarhice, pentru procese industriale.

Echipa acestui proiect este formată din:

- Lappeenranta University of Technology, Finland - coordonator;
- ICI - Romania;
- MSU - Moscow State University Computational Mathematics and Cybernetics Department, Russia;
- ISTC - Information Society Technologies Center, Armenia;
- GIC - V.M.Glushkov Institute of Cybernetics of the Ukrainian Academy of Sciences;
- CIT - Parallel Computing and Architectures Department, Research-Engineering Center of Informational Technologies (RECIT), National Academy of Sciences of Belarus.

Organizat în patru capitole, acest articol urmăreşte structura Raportului tehnic, prezentând, în primul capitol, aspecte legate de definiţia termenilor *data mining* şi *knowledge discovery* pentru ca, în capitolul al doilea, să se urmărească prezentarea caracteristicilor datelor geospaţiale, care trebuie luate în considerare în contextul *data mining*, în general, şi în contextul obiectivelor acestui proiect, în special. Următorul capitol este dedicat sistemelor software pentru *geospatial data mining* (GeoMiner, ADaM, SPIN!) relevante. Ultimul capitol prezintă câteva concluzii privind impactul asupra ştiinţelor legate de informaţia geografică şi aspectele critice al cercetărilor în domeniu. Acest articol, ca de altfel întregul Raport Tehnic, se bazează pe o listă bogată şi actualizată de referinţe bibliografice.

Raportul Tehnic a fost executat ca suport pentru următoarele activităţi ale proiectului INTAS nr. 397:

- tehnici de distilare a datelor pentru masive de date şi modelare predictivă (Task 1);

- activităţi dinamice în data mining (Task 4):
  - prognoze agricole pe baza factorilor meteorologici şi a geomonitorizării regiunilor;
  - modelarea curgerii râurilor folosind reţele neuronale.
- elaborarea de modele de calcul adecvate pentru implementarea clasificării deterministe(Task 7)

şi constituie contribuţia ICI la task-urile menţionate mai sus.

Sugerăm completarea conţinutului acestui Raport Tehnic cu conţinutul unui alt Raport Tehnic, efectuat tot de ICI în aprilie 2002, în cadrul activităţilor aceluiaşi proiect, dedicat "*Tehnicilor şi algoritmilor specifici domeniului geospatial data mining*" [35] şi cu rezultatele prezentate de drnd. mat. Laurenţiu Leuştean în două articole:

- **Liquid Flow Time Series Prediction using Feed-Forward Neural Networks and Rprop Leraning Algorithm**, publicat în "Studies in Informatics and Control", December 2001, pp. 287-300;
- **Liquid Flow Time Series Prediction using Feed-Forward Neural Networks and SuperSAB Learning Algorithm**, articol selectat pentru CONTI'2002 The 5-th INTERNATIONAL CONFERENCE ON TECHNICAL INFORMATICS 18 - 19 October 2002, TIMISOARA, ROMANIA (în curs de publicare).

1. Introducere

Una dintre definiţiile acceptate pe scară largă a termenilor data mining1 şi knowledge discovery este dată de Fayyad et al. [14]:"Data mining/knowledge discovery reprezintă procesul netrivial de identificare de şabloane de date inteligibile, valide, noi, potenţial utilizabile".

Tehnologia Knowledge Discovery (KD) contribuie serios la dezvoltarea următoarei generaţii de sisteme informatice şi sisteme de gestiune de baze de date, prin posibilităţile ei de a extrage informaţie nouă, înglobtă în baze de date eterogene de volum mare, şi de a construi cunoştinţe. Un proces de knowledge discovery include "*selectare de date din depozite mari de date, curăţare, preprocesare, transformare şi reducere, data mining, selecţie (sau combinare) de model, evaluare şi interpreatre şi consolidarea şi utilizarea cunoştinţelor extrase*" [15]. În particular, *data mining* se ocupă cu dezvoltarea algoritmilor de extragere de noi şabloane din statistici, modelare cu reţele neuronale şi vizualizare pentru clasificarea datelor şi identificare de şabloane. *Knowledge discovery* are ca scop crearea mecanismelor prin care informaţia este transformată în cunoştinţe prin intermediul ipotezelor de testare şi al formalismelor teoretoce.

Din definiţia termenului *data mining*, pot fi observate următoarele aspecte:

1. *data mining* nu este o analiză simplă şi nici nu este, în mod necesar, egală cu machine learning; nu este trivial de menţionat că seturile de date considerate sunt mari; în cele mai multe cazuri, este posibilă o analiză statiscă exhaustivă şi este de dorit ca ea să fie cât mai riguroasă (multe metode din *data mining* conţin un grad de nedeterminare, care le permite scalarea la seturi de date masive);
2. unul dintre aspectele necunoscute la începutul procesului şi care trebuie găsit, constă în faptul că *data mining* nu se aplică în cazul în care rezultatul este deja cunoscut, adică în cazul problemelor deterministe sau deductive. *data mining*, în sens larg, ar putea fi o activitate *abductivă* (numită *hypothesis* de filiosoful şi logicianul C.S. Peirce (1878)) care nu acoperă simultan o structură oarecare în cadrul datelor şi o ipoteză de explicare a ei; aceasta va necesita structuri conceptuale sofisticate, prin care o ipoteză poate fi reprezentată într-o maşină; în cadrul *knowledge discovery*, accentul pare să se pună pe metode inductive de învăţare, unde scopul este acela de a construi un model pentru intensitatea (intensitatea se referă în acest caz, mai degrabă, la descrierea generală a unei categorii decât la exemplele ei specifice) unei categorii oarecare din exemplele de instruire; deoarece structura este cunoscută în sens larg, aceasta nu se referă la acţiunea de *mining per se* ci, mai degrabă, la o formă de *knowledge discovery;* o singură excepţie apare atunci când exemplele de instruire ele însele sunt doar o ipoteză, mai degrabă generată din date, decât *a priori*, într-un efort de stabilire a claselor cu care se reprezintă datele, aşa cum este cazul unor instrumente ca AutoClass [10];
3. structura neacoperită trebuie să fie validă, adică trebuie arătat că poate fi o inferenţă semnificativă sau fezabilă cu un grad oarecare de confidenţialitate; metricile de fiabilitate sunt cerute ca suport al ipotezelor prezentate şi pentru a diferenţia semnificativul din marginal sau irelevant;
4. constatările pot fi de domeniul noului; maşina nu are o imagine asupra a ceea ce este cunoscut sau nu de către experţi, adică nu are metode de a mapa noutatea pe domeniul discursului; prin urmare, este posibil să

1 Acest termen ar putea fi tradus ca "*mineritul datelor*". Deoarece până la data publicării acestui articol nu există consens asupra termenului în limba română, pentru a evita confuziile, îl vom utiliza pe cel în limba engleză.

se postproceseze rezultatele astfel încât majoritatea inferenţelor similare să fie grupate laolaltă într-o formă generalizată numită meta-instruire [4];

5. structura neacoperită trebuie să fie utilă, adică să fie explicabilă şi aplicabilă într-o manieră care are sens în contextul domeniul de aplicare curent; seturile mari de date pot să conţină foarte multă structură care ea însăşi nu este utilă şi orientarea efortului pe aceste părţi care sunt intersante este problematică deoarece este, prin definiţie, necunoscută la început.

De obicei, *data mining* se referă la cazurile în care datele sunt prea mari sau extrem de complexe pentru a permite fie o analiza manuală, fie o analiză prin metode de interogare simple. *Data mining* constă din două etape principale:

- preprocesarea datelor, în timpul căreia caracteristicile relevante de nivel înalt sau atributele sunt extrase din date de nivel scăzut şi
- *pattern recognition*, în care recunoaşterea unui şablon de date se face prin intermediul acestor caracteristici (figura 1).

Preprocesarea datelor este, de cele mai multe, ori consumatoare de timp, prin urmare este critică. Pentru a asigura succesul în procesarea din *data mining* este important ca toate caracteristicile extrase din date să fie relevante pentru problemă şi reprezentative ca date:

- în funcţie de tipul datelor care trebuie "minerite", pasul de preprocesare constă în mai multe acţiuni; dacă dimensiunea datelor este foarte mare, va trebui să se recurgă la simplificare şi să se lucreze cu foarte puţine instanţieri sau să se utilizeze tehnici multirezoluţie şi să se lucreze cu datele la o rezoluţie macrogranulară. apoi zgomotul este înlăturat pe cât posibil şi sunt extrase caracteristicile relevante; în unele cazuri, în care sunt disponibile date din diferite surse sau senzori, este necesară fuziunea datelor pentru a permite exploatarea tuturor datelor disponibile pentru problemă; la sfârşitul acestui prim pas, se obţine un vector caractersitic pentru fiecare instanţiere; în funcţie de problemă şi de date, s-ar putea să fie necesară reducerea numărului de caracteristici, folosind tehnici de selecţie sau de reducere a dimensiunii cum ar fi analiza componentei principale [37] sau a versiunilor sale nonlineare; după această preprocesare, datele sunt gata pentru detectarea de şabloane prin intermediul algoritmilor de clasificare, clustering, regresie etc.; aceste şabloane îi sunt afişate utilizatorului în vederea validării; *data mining* este un proces iterativ şi interactiv; ieşirea din oricare dintre paşi sau feedback-ul de la experţi poate să conducă la o rafinare iterativă a unora sau a tuturor acestor activităţi.

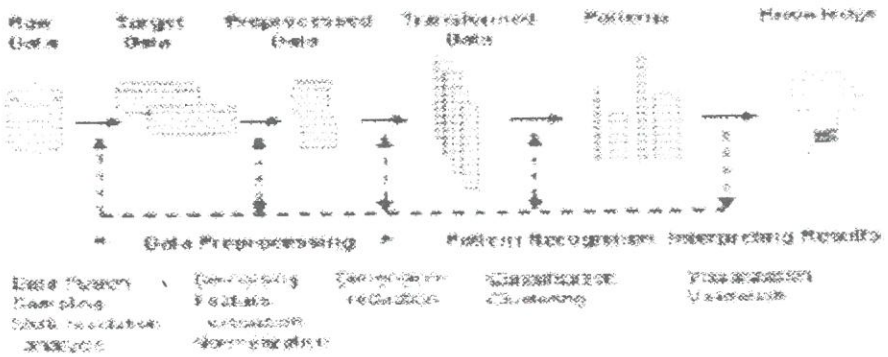

Figura 1: Data mining – un proces iterativ şi interactiv
(adoptată de la Erick Cantu-Paz and Chandrika Kamath, *"On the use of evolutionary algorithms in data mining"*)

Aşa cum menţionează Erick Cantu-Paz şi Chandrika Kamath, în *"On the use of evolutionary algorithms in data mining"* există unele dezbateri legate de definirea termenului de *data mining*, în care majoritatea practicienilor şi a celor care au propus definiţii au căzut de acord că *data mining* este un domeniu multidisciplinar, împrumutând idei de la *machine learning* şi inteligenţă artificială, statistică, calcul de înaltă performanţă, procesare de imagine, optimizare, *pattern recognition* etc. Ca element de noutate, trebuie remarcată confluenţa ramurilor mature ale acestor tehnologii, la un moment dat, în care ele trebuie exploatate în analiza masivelor de date.

În [7] este prezentată în rezumat intersecţia dintre comunităţile academice şi activităţile de *knowledge discovery*. Acest rezumat (Tabelul 1) nu este unul exhaustiv, el prezentând doar câteva dintre iniţiativele şi direcţiile de cercetare cheie.

2. Cunoaşterea geografică

Geografia este o disciplină integratoare: datele necesare abordării ei acoperă un spectru larg de domenii, de interese, de la aspectele sociale până la cele fizice. În continuare, sunt discutate problemele care provin dintr-un amalgam de perspective şi în strânsă legătură cu o infrastructură adecvată dobândirii de informaţii.

- **Complexitatea asociată cu volumul datelor**

Ca toate disciplinele în care este implicat *data mining*, geografia este foarte bogată în date. Multe dintre bazele de date de tip consumator (de ex: care conţin aspecte medicale, tranzacţii financiare), care sunt acum în construcţie, conţin atribute spaţiale şi temporale, oferind, astfel, posibilitatea descoperirii sau confirmării cunoştinţelor geografice [48]. Există acum seturi de date geografice de ordinul terabyte.

Tabelul 1: Rezumatul punctelor de vedere ale diferitelor discipline asupra *data mining* şi *knowledge discovery* (preluare din Buttenfield, B., Gahegan, M., Miller, H., Yuan, M.)

	Baze de date	Statistică	Inteligenţă Artificială	Vizualizare
Regăsire	Reguli de asociere	Şabloane de analiză locală şi teste de inferenţiere globale	Reţele neuronale, arbori de decizie	Vizualizare exploratorie, Visual data mining
Raportare	Regula de listare	Semnificaţie şi putere	Estimare probabilită, câştigare de informaţii	Un stimul în domeniul vizual
Reprezentare	Schema de actualizare, metadate	Modele de potrivire statist., locale sau globale	Grafuri conceptuale, metamodele	Partajare între scenă şi observator
Validare	Testarea slabei semnificaţii	Teste de semnificaţie	Învăţare urmată de verificare	Testarea subiecţilor umani
Optimizare	Reducerea complexităţii de calcul	Reducerea datelor şi simplificare startificată	Căutare stocastică, metode de gradient	Metode ierarhice şi adaptive

- **Complexităţi asociate domeniului**

Semnale interesante şi relevante sunt, adesea, ascunse în întregime de şabloane puternice, care trebuie mai întâi să fie înlăturate. Multe dintre aspectele complexe îşi au originea în codependenţa spaţială şi temporală, care apare între diferitele varietăţi de scări şi din cauze extrem de diverse [57].

- **Complexităţi datorate variaţiei locale**

Sistemele legate de Pământ sunt conectate intrinsec, ceea ce face dificilă analiza izolată, pe o parte, a unui sistem de la care să se modeleze alte aspecte. Rezultatul va apare în formă statistică.

- **Complexităţi cauzate de colectare şi eşantionare**

Cu toate că datele sunt disponibile în a-şi creşte volumul, cazul cel mai frecvent este acela în care trebuie făcută mai degrabă o (re)sortare pentru a surprinde fenomenele de interes, decât să se treacă direct la măsurători.

- **Dificultăţi în formalizarea domeniului geografic**

Una dintre cele mai mari dificultăţi în ceea ce priveşte activitatea de *knowledge discovery* în cadrul domeniului geografic este însăşi complexitatea domeniului. Există, deşi neacceptat pe scară largă, un model conceptual al geografiei [27], iar modelele în GIS-urile comerciale variază, adesea, în mod fundamental, ceea ce conduce la trei probleme distincte:

- datele sunt, de cele mai multe ori, necomensurate ceea ce face ca ele să nu poată fi comparate sau combinate direct;
- este dificil de aplicat cunoaşterea geografică formală procesului de *knowledge discovery*, deoarece astfel de cunoştinţe nu sunt uşor disponibile;
- când noua cunoştinţă nu poate fi "acoperită", este foarte dificil de reprezentat ceea ce vrea să spună ea.

Alte puncte de vedere, legate de aceste probleme, sunt prezentate de Yuan *et al.* [63] şi Miller and Jiawei, [48]. Puse laolaltă, aceste puncte de vedere conduc la necesitatea ca activităţile de descoperire să fie situate în jurul expertului în domeniu: geograful care este, cel puţin în prezent, cea mai bună sursă de înţelegere şi cunoaştere a domeniului.

În rezumat, exemplele de date geografice includ date ce descriu evoluţia fenomenelor naturale în timp, date specifice ştiinţelor pământului, care descriu fenomene spaţiotemporale în atmosferă, secvenţe de imagini bi- şi tridimensionale ale regiunilor geografice sau date care descriu localizarea indivizilor în spaţiul geografic, în funcţie de timp. De obicei, datele geospaţiale conţin informaţii geometrice sau topologice, care se adaugă informaţiilor despre factorii de mediu sau datelor legate de explorarea resurselor naturale [16], [40]. Astfel de date reprezintă cea mai mare parte din datele care au fost colectate de diverse organizaţii. Sistemele de

observare, cum sunt sateliţii sau radarele, pot genera seturi de date foarte mari. De ex:, se estimează că Sistemul de Observare a Pământului al NASA va transmite 50 GB pe oră.

Datele geospaţiale (statice sau care implică evoluţia în timp) se deosebesc, în mod esenţial, de datele *on-line analytical processing* (OLAP), datele multidimensionale generale, datele relaţionale clasice sau datele tranzacţionale. Aplicând normele spaţiului euclidian, datele geografice dau distanţa informaţională şi topologică [30]. De obicei, domeniul atributelor este real. Seturile de date geospaţiale pot să înregistreze informaţii suplimentare ca de ex:, parametrii mediului, definind astfel spaţiul dimensional al atributelor care pot fi corelate. Suplimentar, datele spaţiotemporale pot să reprezinte schimbările intervenite în timp în geometria obiectelor. Activităţile specifice *data mining*, care pot fi aplicate datelor geospaţiale, depind de date şi de problemă. Problema generală poate fi formulată prin găsirea şi descrierea structurii datelor care mai înainte nu au fost cunoscute, şi nu este memorată explicit în baza de date [40].

Geographical Information Science a devenit un organism de cunoaştere, construit pe baza fundamentelor din domenii ca: vizualizare geografică (*geographic visualisation* = GVis), *geographic information systems* (GIS) şi *knowledge discovery in databases* (KDD). Eforturile depuse în ultimii zece ani atât în cercetarea fundamentală, cât şi în cea aplicativă, specifică acestor domenii, au căutat răspunsuri şi la sporirea disponibilităţii datelor şi la implicarea rapidă a tehnologiilor computerizate. În particular, în contextul dezvoltării de metode inovatoare şi producerii de instrumente asociate, s-a căutat facilitarea de decizii politice, evaluări şi planificări apelând, în acest fel, la o audienţă mai mare din partea utilizatorilor de informaţie geografică.

Ţinta principală în cercetarea legată de GVis a fost rolul instrumentelor de vizualizare interactivă care facilitează identificarea şi interpreatrea de şabloane şi relaţii în cadrul datelor complexe [42]. GVis s-au construit pe o bază cartografică, cu GIS, analiza datelor spaţiale şi a imaginilor, eforturi care se corelează cu eforturile ştiinţifice, legate de vizualizarea informaţiei, şi cu eforturile exploratorii din statistică, legate de analiza datelor. Dezvoltarea KDD coincide cu creşterea exponenţială a volumului de date generate de şi disponibile pentru ştiinţă, guvernare şi industrie, date generate, în particular, în format digital.

Termenul de "*knowledge discovery in databases*" a apărut în 1989, în contextul eforturilor de a face distincţie între aplicarea de algoritmi specifici pentru extragerea de şabloane din date (subproces al *data mining*) şi întregul proces, în care *data mining* este un pas în extragerea de cunoştinţe din aceste şabloane [14]. Din literatura consultată, au reieşit numeroase metode KDD, care diferă în ceea ce priveşte dezvoltarea conceptuală, reflectând dezvoltarea separată din cadrul unor domenii ca sistemele de baze de date, machine learning, statistică şi inteligenţă artificială [3], [11].

În ultimii zece ani, au fost acumulate mari cantităţi de date în GIS, cu scopul amplu de analiză exploratorie folosind metode GVis şi KDD şi instrumente asociate. Multe dintre datele legate de mediu, generate recent, aşa cum s-a întâmplat cu cele de la Earth Observation System - EOS, cu sistemele de monitorizare, staţiile meteorologice, includ georeferenţiere. Aspectele spaţiale ale acestor date sunt, de fapt, unul dintre scopurile principale ale analizei – pentru studierea dispersiei poluanţilor, fragmentarea din silvicultură, monitorizarea amenajării teritoriului şi alte aplicaţii. Observaţiile repetate s-au dovedit critice, în ceea ce priveşte formularea de răspunsuri la cele mai importante întrebări ale ştiinţei mediului – cele legate de procesele specifice. Datele georeferenţiate de mediu au, de obicei, componente spaţiale şi temporale.

În rezumat, comunităţile academice au căzut de acord că dezvoltarea tehnologiilor specifice *data mining* (DM) şi *knowledge discovery* (KD) au deschis drumuri noi în cercetare, în general, şi în cadrul cercetării informaţiei geospaţiale, în particular. Abilitatea de a "mineri" datele presupune că există mecanisme de furnizare şi de acces la date. Chiar dacă serviciile de furnizare sunt pe cale să devină disponibile în medii locale şi distribuite, rămân o mulţime de probleme nerezolvate. O parte a activităţii academice trebuie să se dedice infrastructurii suport pentru *data mining* şi *knowledge discovery*. Mecanismele existente nu sunt proiectate să manipuleze probleme specifice ale informaţiei geospaţiale.

În conformitate cu [7], trei caracteristici ale datelor geospaţiale crează provocări în ceea ce priveşte dezvoltarea unui fundament robust de date. Dezvoltarea unei infrastructuri de date, necesară ca suport pentru GIScience, constituie o ţintă a unei alte iniţiative a University Consortium on Geographic Information Science (UCGIS) (spatial data infrastructure). Accentul nu este pus aici pe infrastructura de date spaţiale în sine, ci pe dezvoltarea *data mining* în cadrul infrastructurii.

a) Depozitele de date geospaţiale tind să fie foarte mari

Aşa cum am menţionat anterior, volumul datelor a fost un factor important în tranziţia mai multor agenţii de la furnizarea de date publice prin intermediul mecanismelor fizice (CD ROM, de ex:) la mecanisme electronice [50]. Mai mult, seturile de date GIS existente sunt, de cele mai multe ori, sparte în componente de tip carcateristică şi de tip atribut, care sunt arhivate în mod convenţional în sisteme hibride de management de date.

Cerinţele algoritmice diferă substanţial de managementul de date relaţionale (atribute) şi pentru management de date topologice (caracteristice) [32]. Procedurile de calcul ale KD pot fi diversificate, dacă sunt pe cale să devină complet operaţionale într-un mediu geospaţial de calcul. Chiar în cazul apariţiei şi dezvoltării de noi modele de date pentru GIS integrat, modelul de date hibrid (caracteristică/atribut) va fi păstrat. În practică, integrarea de cunoştinţe va începe să acopere nu numai modele disparate de date dintr-o singură arhivă, ci şi arhive disparate în sisteme disparate de management de date. În legătură cu acestea se află şi gradul şi diversitatea formatelor de date geografice, care prezintă provocări unice. Revoluţia datelor geografice digitale a creat noi tipuri de formate de date printre care se află şi cele tradiţionale "vector" şi "raster". Depozitele de date geografice includ date prost structurate cum sunt imagini şi *geo-referenced multi-media* [8]. Descoperirea de cunoştinţe geografice din date georeferenţiate multimedia este un aspect mult mai complex al problemei KD pentru date multimedia [64].

b) A doua caracteristică a datelor geospaţiale se leagă de introducerea caracteristicilor datelor colectate ciclic

Descoperirea de date trebuie să se acomodeze ciclurilor de colectare, care pot fi necunoscute (ca de ex: identificarea ciclurilor de schimbare în dislocările geologice majore) sau care se pot schimba de la un ciclu la altul atât în timp, cât şi în spaţiu (de ex: şabloane de dispersie a epidemiilor sau a substanţelor toxice). Integrarea informaţiei din modele multiple de date este recunoscută ca fiind o ţintă a uneia din temele de cercetare ale UCGIS (Achiziţia şi integrarea de date spaţiale), şi nu ne propunem să o dezvoltăm în acest material. În schimb, cercetătorii din domeniul KD pot să acorde atenţie problemelor de raţionament şi modelare pe cicluri mai lungi sau mai scurte de timp. De exemplu, *geospatial knowledge discovery* (GKD) ar putea să suporte traseele unei furtuni în timp real sau predicţia unei avalanşe sau alte evenimente localizate, legate de evoluţia vremii. Aspectele legate de infrastructură, care urmează să fie cercetate includ, de ex:, dezvoltarea de *real-time data mining* şi utilizarea de instrumente KD pentru ghidonarea corelării de şabloane de date descoperite în timp, validarea tendinţelor datelor dealungul discontinuităţilor temporale ş.a. Deoarece înţelegem mai puţin despre natura timpului decât a spaţiului, metodologiile de arhivare de date, în vederea facilitării de căutări spaţiale ciclice, rămân nefinisate. Extensia la care se pot identifica şabloane de date va fi determinată parţial sau în întregime prin organizarea datelor într-o arhivă. Cercetarea privind cea mai bună structură sau pentru reordonarea datelor pentru activităţile specifice de KD nu este, încă, demarată în alte teme de cercetare ale UCGIS.

c) A treia caracterisică de relevanţă a DM/KD se aplică mai degrabă la o caracteristică a datelor fundamentale decât asupra datelor

Peste trei milioane de Website-uri sunt online. În acest caz, cele mai bune motoare de căutare pot să localizeze cel mult o treime din paginile accesibile [49], [50]. Astfel de surse de date include, dar nu se limitează la site-uri publice de date de domeniu colectate în ţările dezvoltate, site-uri de comunităţi de date extrem de localizate cum ar fi site-uri prezentând vecinătăţile unor oraşe sau ale comunităţii active dintr-un teritoriu şi surse similare de date necunoscute sau cunoscute prin convenţii în cadrul infrastructurii de date geospaţiale. Acest tip de KD se ocupă de întregul Internet ca un depozit foarte mare, descentralizat de date şi furnizează contribuţii la o infrastructură globală. Este paradoxal faptul că o cantitate din ce în ce mai mare de date devine disponibilă prin Internet, în timp ce datele sunt din ce în ce mai greu de localizat, regăsit şi analizat. Aceasta se întâmplă şi datorită lipsei din Internet a unui catalog inteligibil sau index [6]. Fără o infrastructură coordonatoare, multe dintre sursele de date şi serviciile disponibile astăzi rămân în esenţă inaccesibile.

Datele geografice au şi alte proprietăţi unice, care reclamă atenţie şi tehnici speciale, existente în domeniul măsurătorilor geografice. În timp ce alte aplicaţii KD implică spaţii superior dimensionate, datele geografice sunt unice întrucât până la patru dimensiuni ale spaţiului informaţional sunt interrelaţionate şi furnizează contextul de măsurare pentru dimensiunile rămase. Cadrul de măsurare cel mai des adoptat este topologia şi geometria asociată cu spaţiul euclidian. Cu toate acestea, unele fenomene geografice au proprietăţi neeuclidiene. Exemple de acest fel sunt timpii de călătorie în zonele urbane, imaginile mentale ale spaţiului geografic şi propagarea bolilor în spaţiu şi timp [45]. Proiecţia datelor spaţiale, în contexte de măsurare potrivite, poate să ajute căutarea de şabloane în *geographic data mining*. Informaţia inerentă contextului de măsurare geografic este, adesea, ignorată de instrumentele de inducţie şi machine learning [18].

Adesea, atributele geografice măsurate arată proprietăţile dependenţei şi eterogenităţii spaţiale. Primele se referă la tendinţa atributelor de a fi legate de unele locaţii din spaţiu (de obicei, acestea sunt vecinătăţile). Ultimele se referă la nestaţionaritatea majorităţii proceselor geografice, înţelegându-se că parametrii globali nu reflectă foarte bine procesul care apare la o locaţie particulară. În timp ce aceste proprietăţi au fost tratate ca neplăceri, cercetarea contemporană, ajutată de progresele din domeniul tehnologiei informaţiei geografice, furnizează instrumente care pot să explice aceste proprietăţi pentru înţelegerea fenomenelor geografice [1], [5], [17], [25]. Unele cercetări în GKD sugerează că, ignorând aceste proprietăţi, sunt afectate şabloanele derivate din tehnicile *data mining*. Este necesar un efort de cercetare mai mare, în ceea ce priveşte tehnicile scalabile de capturare a dependenţei şi eterogenităţii spaţiale în GKD.

Un al treilea aspect unic al informaţiei geografice în KD este complexitatea obiectelor şi şabloanelor geospaţiale. În majoritatea domeniilor negeografice, obiectele de date pot fi foarte bine reprezentate discret în cadrul spaţiului informaţional fără a pierde proprietăţi importante. Nu este cazul obiectelor geografice: dimensiunea, forma şi graniţele pot să afecteze procesele geografice, înţelegând că obiectele spaţiale nu pot fi

reduse la puncte sau caracteristitici lineare simple, fără a se pierde informaţii. Relaţii cum sunt distanţa, direcţia şi conectivitatea sunt mult mai complexe cu obiectele dimensionale [12], [51], [55]. Transformările în timp, între aceste obiecte, sunt complexe şi purtătoare de informaţii [34]. Scara şi granularitatea, în ceea ce priveşte măsurarea timpului, pot fi complexe, făcând ca, o simplă actualizare de dimensiune a spaţiului, să includă timpul [35], [57]. Dezvoltarea de instrumente scalabile pentru extragerea de şabloane din colecţii de obiecte spaţiotemporale diferite este una dintre provocările majore. Astfel, deşi complexitatea şabloanelor şi a regulilor spaţiotemporale poate fi descurajantă, capătă sens o altă provocare derivată probabil prin intermediul "*meta-mining*" [37].

3. Sisteme software pentru domeniul Geospatial Data Mining/Knowledge Discovery

În "*A Data Miner's Tools*", din BYTE/October 1995, Karen Watterson [60] explică trei categorii de software dedicate domeniului *data mining*. *Instrumentele interogare-şi-raportare*, în forma lor simplificată şi uşor de utilizat, cer asistare umană şi legături în bazele de date sau alte formate speciale. *Instrumentele de analiză multidimensională* (multidimensional analysis = MDA) cer mai puţină intervenţie umană, dar au nevoie de date în formate speciale. *Agenţii inteligenţi* sunt virtual autonomi, sunt capabili să-şi facă propriile observaţii şi să tragă concluzii pe care pot să le manipuleze ca forme libere în paragrafe de text.

Data Mining aşa cum este definit în primul capitol, foloseşte *machine learning*, tehnici statistice şi de vizualizare pentru a descoperi şi prezenta cunoştinţe într-o formă uşor inteligibilă de către factorii umani.

Explozia în datele georeferenţiate, prilejuită de dezvoltarea tehnologiei informaţiei (Information Technology = IT), cartografiei digitale, teledetecţiei şi răspândirii pe scară largă a GIS, accentuează importanţa dezvoltării metodelor inductive, conduse de date în analiză geografică şi modelare pentru a facilita crearea de noi cunoştinţe şi pentru a ajuta procesul descoperirii ştiinţifice. În prezent, există un număr de instrumente pentru *data mining*, care sunt proiectate să asiste procesul de explorare a unor cantităţi mari de date în căutarea de şabloane recurente şi relaţii (de ex: STATISTICA, Clementine, MineSet, Intelligent Data Miner etc.). Aceste pachete au fost dezvoltate, în principal, în scopul de a analiza baze de date comerciale foarte mari, pentru a modela şi a face predicţii în ceea ce priveşte comportamentul consumatorilor. Accentuarea pe predicţie poate să le limiteze utilitatea în ceea ce priveşte GIS, unde *spatial pattern recognition* poate fi o activitate mult mai utilă. În contextul GIS, există trei argumente care pot fi invocate în favoarea dezvoltării şi aplicaţiilor instrumentelor pentru *data mining* pentru a exploata invazia geoinformaţiilor. Mulţi dintre specialiştii implicaţi în *data mining* cred că instrumentele lor vor lucra pe oricare şi pe toate datele indiferent de subiectul de origine. *Geographical Data Mining* trebuie privit ca un tip special de *data mining* (aşa cum este prezentat şi în capitolul al doilea), care ştie să proceseze funcţii generice similare ca şi instrumentele convenţionale de *data mining*, dar modificate sau construite special pentru a ţine cont de caracteristicile geoinformaţiilor, de stilurile diferite şi de cerinţe de analiză şi modelare, relevante pentru lumea GIS, dar şi de natura specifică a explorării geografice.

3.1. GeoMiner: Un sistem pentru descoperirea de cunoştinţe pentru baze de date spaţiale şi Geographic Information Systems

Prezentare generală

GeoMiner este un sistem pentru descoperire de cunoştinţe pentru baze de date spaţiale, dezvoltat în Database Systems Research Laboratory, School of Computing Science, Simon Fraser University. Această secţiune se bazează pe prezentarea publică de la adresa: http://db.cs.sfu.ca/DBMiner.

Spatial data mining constă în a "mineri" informaţia spaţială de nivel superior şi cunoştinţele din baze de date spaţiale mari. Un prototip de sistem pentru *spatial data mining*, GeoMiner, a fost proiectat şi dezvoltat pe baza experienţei în cercetarea şi dezvoltarea unui sistem relational data mining, DBMiner, şi pe baza cercetărilor din domeniul spatial data mining. Puterea lui GeoMiner în *data mining* include "mineritul" a trei tipuri de reguli: reguli caracteristice, reguli de comparare şi reguli de asociere în baze de date geospaţiale cu o extensie planificată pentru a include "mineritul" regulilor de clasificare şi clustering. GeoMiner include modulul de construcţie spatial data cube, modulul spatial on-line analytical processing (OLAP), precum şi module de *spatial data mining*. De asemenea, a fost proiectat şi implementat un limbaj de *spatial data mining language*, GMQL (Geo-Mining Query Language), ca o extensie a Spatial SQL, pentru spatial data mining. A fost construită şi o interfaţă interactivă şi prietenoasă şi au fost implementate instrumente de vizualizare a cunoştinţelor spaţiale descoperite.

Descrierea proiectului

GeoMiner permite găsirea de cunoştinţe intersante din baze de date spaţiale mari. Metodele de *spatial data mining* au fost aplicate pentru a extrage astfel de cunoştinţe din baze de date spaţiale mari. O dată cu progresul făcut de cercetare în *data mining* şi *data warehousing* din ultimii ani, au fost dezvoltate multe sisteme de *data mining* şi *data warehousing* pentru "mineritul" de cunoştinţe în baze de date relaţionale şi *data warehouses*.

Spatial data mining este un subdomeniu care se ocupă cu extragerea de cunoştinţe implicite, relaţii spaţiale sau de alte şabloane interesante, memorate neexplicit în baze de date spaţiale. Cu o mare cantitate de date spaţiale, colectate de sistemele satelitare, sistemele de teledetecţie, sistemele de vânzări regionale şi de alte instrumente de colectare de date, este inevitabilă dezvoltarea de instrumente pentru descoperirea de cunoştinţe interesante din bazele de date spaţiale mari. În plus, multe dintre bazele de date relaţionale conţin şi ele informaţii spaţiale cum sunt adrese ale reşedinţelor clienţilor sau localizarea unui magazin/depozit. Aceste adrese pot fi geocodificate prin extragerea coordonatelor spaţiale, care pot fi memorate într-o bază de date spaţială împreună cu alte date spaţiale. Este important să se "mineze" cunoştinţe legate atât de obiecte spaţiale, cât şi de obiecte nespaţiale în bazele mari de date. În studiile făcute până acum, din păcate, nu sunt cunoscute multe sisteme de *spatial data mining*.

Progresele recente în ceea ce priveşte structurile datelor spaţiale şi ale bazelor de date spaţiale fac posibilă crearea de baze de date spaţiale mari, care pot fi efectiv interogate. Aceste progrese în combinaţie cu cercetările din cadrul raţionamentului spaţial, precum şi cu cele din domeniul *data mining* în baze de date relaţionale, promovează cercetările din cadrul *spatial data mining*.

Grupul de cercetare GeoMiner din cadrul Intelligent Database Systems Research Laboratory a lucrat mulţi ani în domeniul *data mining* şi, în special, în *spatial data mining*. Sistemul GeoMiner include modulul de construcţie *spatial data cube*, modulul *spatial on-line analytical processing* (OLAP), precum şi module de *spatial data mining* pentru "mineritul" de reguli caracteristice, reguli de comparare, reguli de clasificare, reguli de asociere şi clustering.

Arhitectura SAND (Spatial And Nonspatial Data) se aplică la modelarea bazelor de date spaţiale, iar modulele de *spatial data mining* includ atât "mineritul" cunoştinţelor spaţiale, cât şi relaţiile între componentele spaţiale şi nespaţiale. De asemenea, a fost proiectat şi implementat un limbaj de spatial data mining, GMQL (Geo-Mining Query Language), ca o extensie a Spatial SQL. A fost construită şi o interfaţă interactivă şi prietenoasă şi au fost implementate instrumente de vizualizare a cunoştinţelor spaţiale descoperite.

Architectura

Sistemul GeoMiner este o extensie şi o evoluţie de la un sistem relaţional pentru *data mining* DBMiner, cercetat şi dezvoltat de acelaşi laborator. Sistemul DBMiner (a se vedea referinţele şi sistemul prin intermediul http://db.cs.sfu.ca/DBMiner) conţine următoarele cinci module funcţionale: caracterizator, comparator, asociator, predictor şi clasificator. Numeroase alte module se află, încă, în faza de cercetare şi dezvoltare. DBMiner este implementat prin integrarea tehnicilor de *data mining* şi *data warehousing*, incluzând *data cube construction and manipulation, attribute-oriented induction, multi-level association analysis, statistical data analysis, machine learning* etc. pentru "mineritul" datelor relaţionale.

GeoMiner este construit pe baza sistemului DBMiner. Funcţiile pentru operaţiile OLAP nespaţiale şi pentru "mineritul" datelor nespaţiale sunt direcţionate către sistemul DBMiner; în timp ce funcţiile pentru "mineritul" datelor spaţiale şi pentru relaţiile între datele spaţiale şi nespaţiale sunt procesate de funcţii dedicate ale GeoMiner.

Principalele funcţii ale sistemului includ "mineritul" a cinci tipuri de reguli de cunoştinţe în baze de date spaţiale, integrarea de tehnologii pentru data mining şi data warehousing, "mineritul" interactiv de reguli multinivel, integrarea cu baze de date relaţionale comerciale şi GIS, precum şi multe forme de ieşire, incluzând hărţi generalizate, relaţii generalizate, reguli multinivel etc.

Arhitectura generală a GeoMiner constă din:

1. o interfaţă grafică cu utilizatorul pentru "minerit" interactiv şi afişare de rezultate ale data mining sub formă de tabele, hărţi etc.;
2. un set de module de descoperire, incluzând cele cinci module existente: geo-caracterizatorul, geo-comparatorul, geo-clasificatorul, analizorul de geo-cluster şi geo-asociatorul, precum şi alte două module în dezvoltare: geo-predictorul şi analizorul de geo-pattern;
3. un server de baze de date spaţiale, care include MapInfo Professional 4.1;
4. *data cube mining engine* care se bazează pe nucleul de descoperire DBMiner pentru manipulare de date multidimensionale;
5. baza de date şi baza de cunoştinţe care memorează datele spaţiale şi nonspaţiale şi
6. ierarhiile conceptuale.

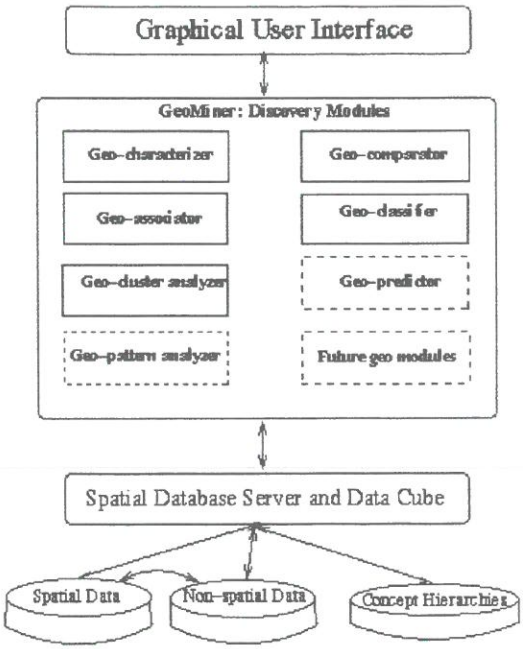

Figura 2: Architectura GeoMiner (preluată de la http://db.cs.sfu.ca/DBMiner.)

Modulele funcţionale importante

Geo-caracterizatorul

Acest modul "minereşte" un set de reguli caracteristice la niveluri multiple de abstractizare, de la un set de date relevant dintr-o bază de date spaţiale. El furnizează utilizatorilor un punct de vedere multinivel, multiunghiular asupra datelor dintr-o bază de date spaţială.

Geo-comparatorul

Acest modul "minereşte" un set de reguli de comparare, care sunt în contrast cu caracteristicile generale ale diferitelor clase ale diferitelor seturi de date, dintr-o bază de date. Se compară un set de date, cunoscut ca şi clasă ţintă, cu unul sau mai multe seturi de date, cunoscute ca şi clase contrastante.

Geo-asociatorul

Acest modul găseşte un set de reguli de asociere spaţial relaţionate din unul sau mai multe seturi de date dintr-o bază relaţională. O regulă de asociere arată apariţia curentă de şabloane (sau relaţii) dintr-un set de articole de date dintr-o bază de date. O regulă de asociere spaţială tipică este de forma ``A -> B (s%, c%)" unde A şi B sunt seturi de predicate spaţiale sau nonspaţiale, s este suportul regulii (probabilitatea ca A şi B să fie păstrate amândouă în toate cazurile posibile), iar c este gradul de confidenţialitate al regulii (probabilitatea condiţională ca B să fie adevărat sub condiţia A).

Analizorul de Geo-cluster

Acest modul găseşte clustere de puncte cu descrieri nespaţiale relevante. El foloseşte un algoritm eficient, CLARANS, pentru a procesa spatial clustering.

Prin folosirea inducţiei orientate-atribut se găsesc descrierile nespaţiale ale clusterelor.

Geo-classificatorul

Acest modul adoptă o metodă de inducţie generalizată, bazată pe arbori de decizie, pentru a construi un arbore de decizie care clasifică setul datelor relevante în conformitate cu unul dintre atributele nespaţiale. Se afişează arborele de clasificare şi, prin poziţionarea pe unul dintre noduri, utilizatorul poate să pună în evidenţă regiunile de hartă corespunzătoare.

Evoluţia GeoMiner

Echipa de la Database Systems Research Laboratory, School of Computing Science, Simon Fraser University îşi propune să continue cercetarea şi dezvoltarea în următoarele direcţii:

- sporirea puterii şi eficienţei mecanismelor de descoperire, incluzând îmbunătăţirea calităţii regulilor şi performanţei sistemului pentru modulele funcţionale existente;
- nu cu mult timp în urmă, echipa s-a concentrat pe descoperirea de cunoştinţe doar dintr-o singură hartă tematică; intenţia sa este de a aborda algoritmi eficienţi, care să permită manipularea de hărţi tematice multiple;
- versiunea actuală de clasificator construieşte arborele de decizie, bazându-se pe proprietăţi nespaţiale ale obiectelor, iar, la sfârşit, este procesată vizualizarea spaţială datelor; echipa şi-a propus să dezvolte noi algoritmi care pot să ţină cont şi de proprietăţile spaţiale ale obiectelor clasificate;
- în plus faţă de modulele menţionate anterior, echipa va implementa un geo-predictor; câteva dintre aspectele realţionale ale predictorului au fost deja implementate în sistemul DBMiner; de asemenea, urmează să fie încorporată şi o parte de analiză de şablon în cadrul sistemului GeoMiner;
- este planificată şi proiectarea şi îmbunătăţirea interfeţelor de nivel înalt, prietenoase pentru *interactiv spatial data mining* şi *visual presentation* pentru cunoştinţele descoperite.

3.2. Data Mining System Toolkit for Earth Science Data

Information Technology and Systems Center (ITSC) de la Universitatea din Huntsville, Alabama, a dezvoltat un sistem pentru *data mining*, care permite cercetătorilor şi altor utilizatori să culeagă informaţii şi cunoştinţe din munţii de date din Ştiinţele Pământului. Seturile de date din cadrul Ştiinţelor Pământului şi, în general, datele spaţiale sunt variabile ca formate, scări şi rezoluţie. Prin cercetările desfăşurate în domeniile *data mining, coincidence data searching, object - relational databases, Open GIS, data integration şi interoperable distributed data systems*, ITSC a produs o varietate de sisteme software, care ajută la a face seturile de date universal accesibile, accesibile şi utile. Prin angajarea construcţiilor orientate-obiect şi a instrumentelor de dezvoltare ITSC s-a creat situaţia de a utiliza structura sistemului pentru a furniza şi alte funcţionalităţi cum ar fi subsetare de date generice şi generarea de date la cerere.

Această secţiune se bazează pe articolul: "*Data Mining System Toolkit for Earth Science Data*", scris de Ken Keiser, John Rushing, Helen Conover, Sara Graves, care este disponibil la adresa: http://webtech.ceos.org/eogeo99/Papaers/Keiser/Adam_EOGEO.html.

Sistemul Algorithm Development and Mining (ADaM) dezvoltat la ITSC a fost instrumentul în detectarea fenomenelor, extragerea de caracteristici şi furnizarea de instrumente pentru analiză de date şi procesare, care integrează variabile spaţiale şi temporale din seturile de date specifice Ştiinţelor Pământului. Acest sistem constă dintr-un cadru de bază pentru interschimbul de date şi un set de module *plug in*, care lucrează în acest context. Modulele includ filtre de intrare, module de analiză şi filtre de ieşire. Filtrele de intrare/ieşire translatează seturile de date specifice într-o zonă de tipuri de date şi formate la şi de la o reprezentare internă puternică şi flexibilă. În cadrul sistemului de mining, datele sunt administrate într-o structură standard, care permite aliniere spaţială şi temporală pentru integrare şi analiză. Modulele de analiză încărcate dinamic, dezvoltate şi personalizate de către cercetători, manipulează structura internă, sunt reutilizabile, şi nu sunt constrânse de diferenţele din formatele externe. Sistemul ADaM are multe module de *data mining, pattern recognition, image processing, subsetting, gridding*. Acest sistem are, de asemenea, filtre pentru o varietate de formate specifice, precum şi filtre noi şi pot fi adăugate cu uşurinţă module de analiză. Cercetătorii de la ITSC împreună cu alţi colaboratori utilizează sistemul pentru *data mining* în studiile despre clasificarea de textură, procesare de imagini şi analiză statistică pentru aplicaţii cum sunt managementul în silvicultură şi modelarea atmosferică.

Contextul de lucru ADaM

Iniţial, ITSC a dezvoltat sistemul ADaM în cadrul unui grant de la NASA pentru a investiga noi metode de procesare a volumelor mari de date de la Earth Observing System (EOS) şi teledetecţie. Scopul grantului a fost acela de a face căutări pe valori de date, precum şi pe metadate, şi de a îmbunătăţi metadatele limitate, disponibile, de obicei, pentru Ştiinţele Pământului prin catalogarea informaţiilor bazate pe conţinutul datelor. Ca subobiectiv, s-a urmărit adăugarea de algoritmi care să poată suporta detectarea unei varietăţi de fenomene geofizice în datele "minerite". ITSC a dezvoltat şi procesele de generare şi de memorare de metadate bazate pe conţinut, care pot fi regăsite de cercetători prin intermediul interfeţelor bazate pe web, care îi conduc la seturile actuale de date cerute în numeroase analize şi studii [33]. Acest sistem pentru *data mining* a fost utilizat şi pentru alte studii care se refereau la clasificare de textură, procesare de imagini şi analiză statistică.

Mediul ADaM

Motorul ADaM este proiectat pentru a extrage conţinut bazat pe metadate din arhivele dedicate Ştiinţelor Pământului [33]. Acesta poate detecta fenomene sau evenimente care sunt de interes pentru oamenii de ştiinţă şi poate memora informaţia într-un mod care facilitează procesul de căutare şi de ordonare a datelor. Unele rezultate de acest tip sunt memorate în Eureka, un motor de căutare de date spaţiale, coincident folosit pentru a găsi coincidenţe între fenomenele *mining-generated*, evenimentele climatologice şi informaţii statistice, cum ar fi graniţele de judeţ sau ale bazinelor hidrologice. Acest motor furnizează şi alte capabilităţi de ordonare ca de ex: generarea de date prin intermediul aplicaţiilor client. Procesarea client poate să includă *gridding, resampling*, filtrare, conversie de format sau alte analize depinzând de cerinţele clienţilor. De ex:, ADaM poate să genereze lunar o imagine a totalului pluviometric de la datele provenind de la radar. Atât instrumentele de interogare spaţială, cât şi aplicaţiile client sunt aplicaţii web şi, în felul acesta, clienţii sunt capabili să ruleze aproape în orice mediu. Figura 3 prezintă o arhitectură generală a mediului pentru *data mining*, AdaM.

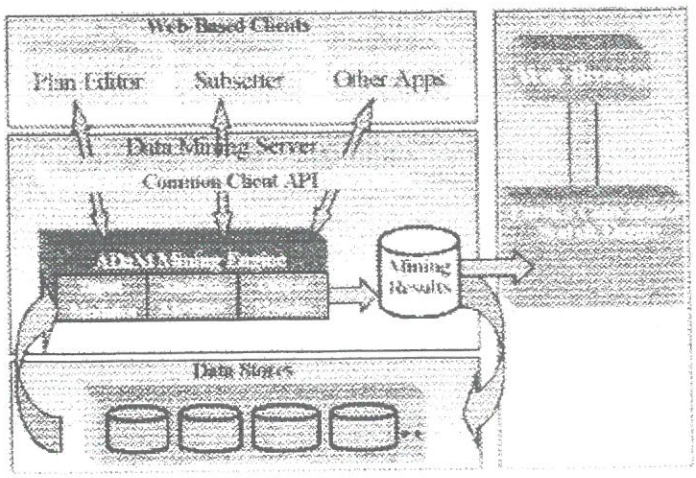

Figura 3: Mediul ADaM pentru data mining (preluată din prezentarea "*Data Mining System Toolkit for Earth Science Data*", făcută de Ken Keiser, John Rushing, Helen Conover, Sara Graves şi disponibilă la adresa: http://webtech.ceos.org/eogeo99/Papaers/Keiser/Adam_EOGEO.html)

Arhitectura de procesare ADaM

Specificul datelor spaţiale constă în faptul că tind să înţeleagă mai multe forme şi dimensiuni în termeni de formate, scări şi rezoluţii. Acestea s-au dovedit a fi probleme consumatoare de timp atunci când cercetătorii se aşteptau să integreze seturi noi de date în studiile şi modelele lor. În ideea depăşirii acestor probleme, ADaM a fost proiectat pentru a manipula intern toate datele într-un mediu comun fără a pierde formatele de intrarea şi de ieşire actuale. Această metodă izolează efectiv modulele de analiză internă de setul de date specific. Se întâmplă, în mod frecvent, ca să fie necesari mai mulţi paşi distincţi pentru îndeplinirea unei activităţi. Este adeseori de dorit ca aceasta să se producă fără scrierea de rezultate intermediare pe disc. Pentru atingerea acestui scop, proiectanţii lui AdaM au adoptat metoda *data pipeline*. Rezultatele provenite dintr-o operaţie de procesare constituie intrare pentru următoarea operaţie. Mai multe operaţii pot fi înlănţuite într-un singur plan de procesare. Figura 4 ilustrează arhitectura funcţională a motorului AdaM.

3.3. Integrarea GVis, GIS şi KDD pentru explorarea datelor spaţio-temporale

Această secţiune prezintă o metodă de abordare a unui proces pe trei niveluri ca o strategie pe termen lung pentru a integra metode GVis şi KDD cu un GIS temporal. Scopul principal este acela de a dezvolta tehnici inovatoare şi instrumente associate pentru explorarea de date (care includ informaţie cantitativă şi calitativă) pentru, a ajunge la şabloane spaţiotemporale valide, noi, potenţial utile şi inteligibile şi pentru a pocesa date. Rezultatul este un proiect de vizualizare bazată pe cunoştinţe pentru GIS temporal, care implică procesul de percepţie cognitivă şi procesarea automată de informaţii cu multe decizii luate de utilizatorii de informaţie geografică (GI) în ceea ce priveşte modul în care se potrivesc modelele sau modul în care se determină şabloanele spaţiotemporale din date.

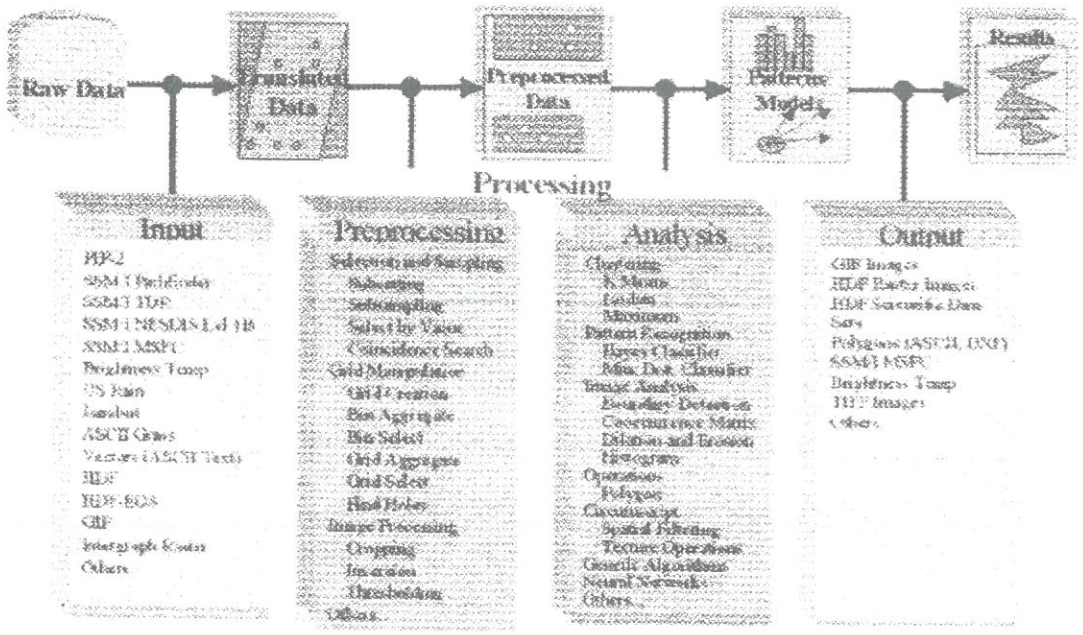

Figure 4: Fluxul de date în cadrul AdaM (preluată de la "*Data Mining System Toolkit for Earth Science Data*", autori: Ken Keiser, John Rushing, Helen Conover, Sara Graves http://webtech.ceos.org/eogeo99/Papaers/Keiser/Adam_EOGEO.html)

Bazându-se pe articolul scris de Monica Wachowicz, "*Integrating GVis, GIS and KDD for Exploring Spatio-Temporal Data*", această secţiune prezintă o discuţie preliminară asupra modului în care integrarea GVis, GIS şi KDD poate naşte noi provocări rezultând din complexitatea procesului de explorare de date şi capabilităţile suplimentare ale sistemului integrat ţintă. Din punct de vedere sistemic, rezultatele sunt legate de suportul efectiv al interacţiunilor utilizator-date atât în modele de date, cât şi în interfeţele de nivel înalt. Din punctual de vedere al utilizatorilor de informaţie geografică, principalul rezultat constă în a face procesul de construire de cunoştinţe cât mai flexibil şi de a facilita explorarea interactivă a datelor spaţiotemporale multivariate. În secţiunile următoare, sunt trecute în revistă principiile şi dezvoltările cheie din ultimi zece ani din domeniile GVis, GIS şi KDD. Această trecere în revistă furnizează o bază de la care se poate explora integrarea potenţială a GVis, GIS şi KDD.

Cunoştinţele câştigate din cercetările desfăşurate în vederea integrării GVis, GIS şi KDD vor fi fructificate în proiectarea viitoarei generaţii de GIS, care va furniza un singur mediu pentru explorarea bazei de date, analizare şi vizualizare. Cu alte cuvinte, această integrare va furniza instrumente pentru regăsirea informaţiei, explorare şi analizare şi va permite utilizatorilor să controleze interactiv vizualizări animate ale conţinutului bazei de date, interogări şi rezultate ale interogărilor. Instrumentele de succes vor fi: *puternice* pentru a furniza imediat tehnologie cu valoare-adăugată, *flexibile* în procesarea de date şi prezentarea ieşirilor pentru a evita conflictele cu clienţii, *sensibile* în a furniza un mediu superior interactiv şi *uşor de utilizat* pentru a încuraja experimentarea spontană.

Analiza exploratorie de date este un proces iterativ, în care interogările de nivel înalt (conceptual) conduc la interogări specifice, ale căror răspunsuri sunt examinate de către utilizatorii de informaţie geografică în vederea găsirii de şabloane interesante care, la rândul lor, pot sugera noi interogări. Provocarea implicată de acest tip de explorare este, în principal, legată de furnizare rapidă de operaţii de analiză, incrementale şi reversibile, care generează continuu bucle de reacţie. Este necesară o interfaţă uilizator grafică pentru a reorganiza instantaneu datele, pentru a juxtapune caracteristici în conformitate cu diverse criterii şi pentru a vizualiza relaţiile între diversele proprietăţi ale acestor caracteristici. O caracteristică poate fi un obiect fizic, obiect abstract sau un eveniment. Integrarea GVis, KDD şi GIS va combina uşurinţa în utilizare a sistemelor de manipulare directă/interactivă cu puterea sistemelor de interogare de baze de date.

Pentru atingerea acestui deziderat, este necesar un proces de management pentru a combina metodele şi instrumentele asociate, care facilitează înţelegere ştiinţifică a seturilor mari de date, folosind un mediu de explorare singular. În această secţiune, este propusă şi discutată o metodă de procesre în trei niveluri pentru executarea acestei integrări la trei niveluri:

- nivelul conceptual pentru înlănţuirea etapelor găsite în GIS, GVis şi KDD,
- nivelul operaţional pentru integrarea metodelor dezvoltate independent în fiecare dintre domenii,
- nivelul de implementare pentru combinarea diferitelor instrumente într-un singur mediu de lucru (sistem).

Principalul rezultat al acestei abordări constă în proiectarea unei vizualizări bazate pe cunoştinţe pentru GIS temporal, în care accentul se mută de la metoda deductivă la cea condusă de date. În loc de a se căuta ipoteze pentru un model care pare că se potriveşte datelor disponibile, accentul este pus acum pe date şi pe descoperirea inerentă de relaţii în cadrul lor. Mai mult, această abordare va suporta metode noi de a interacţiona cu seturi mari de date, va avea flexibilitate în a se ocupa de scări de variaţie a spaţiului şi timpului, mecanisme noi de identificare şi trasare de incertitudine, precum şi versatilitate în ceea ce priveşte manipularea de formate multiple de date.

Nivelul conceptual

La nivelul conceptual, sunt identificate cerinţele utilizatorului translatate în scopuri de nivel înalt, care urmează să fie atinse în construcţia cunoştinţelor folosite în progresul ştiinţific, sporirea profitului în afaceri, administrarea resurselor naturale şi aplicaţii specifice. Aspectele critice ale procesului de construcţie de cunoştinţe de la acest nivel sunt:

- ce tip de date spaţiotemporale se estimează a fi explorate (adică specifice mediului, socio-economice, discrete sau continue),
- ce rezultate particulare sunt aşteptate de la acest proces (adică generarea de ipoteze, predicţia unei schimbări ulterioare),
- cine sunt utilizatorii cunoştinţelor obţinute (adică, oameni de ştiinţă, analişti politici).

Deciziile privind aest nivel conceptual acţionează ca şi constrângeri de integrare ale etapelor GVis, GIS şi KDD (a se vedea tabelul 2 pentru o trecere în revistă a etapelor actuale ale GVis, GIS şi KDD). Abordările particulare pot să identifice cerinţe mai degrabă pentru explorare vizuală de date, decât pentru un algoritm automatizat pentru data mining, conducând la o amalgamare a etapelor GVis (cunoaşterea percepţiei vizuale) şi a etapelor GIS (cunoaşterea datelor). În acest caz, amalgamarea va fi definită în conformitate cu constrângerile din domeniile GVis şi GIS. O constrângere comună constă în faptul că, înainte ca datele să poată fi vizualizate efectiv, pot fi cunoscute diverse metadate referitoare la structură şi tip. Datele sunt organizate în baze de date (fişiere flat, tabele) în care fiecare bază de date se leagă de o metodă particulară de capturare de date.

Tabelul 2: Etapele principale în GVis, GIS şiKDD (preluare din Monica Wachowicz, "*Integrating Gvis, GIS and KDD for Exploring Spatio-Temporal Data*")

GVis *Explorare condusă de percepţie vizuală*	**GIS** *Explorare condusă de instrucţiuni de interogare*	**KDD** *Explorare condusă de proceasrea automată de informaţii*
Explorare	Colectare de date	Selecţia de date
Confirmare	Modelare de date	Pre-Procesare
Sinteză	Manipulare de date	Data Mining
Prezentare	Ieşire/Prezentare	Interpretare/Evaluare

Nivelul operaţional

Nivelul operaţional se ocupă cu specificarea şi combinarea celor mai potrivite metode, în vederea atingerii scopurilor conceptuale. Această integrare a metodelor GVis, GIS şi KDD este foarte importantă în obţinerea de avantaje de la experţii umani, specialiştii în analiză (expertiza de domeniu şi abilităţi de visual pattern recognition) şi de la computere (putere mare de procesare). În continuare, se propune o perspectivă de integrare bazată pe metodele GVis, GIS şi KDD pentru explorarea ştiinţifică a datelor spaţio-temporale (tabelul 3), bazată pe o metodologie ***task analysis*** (Kirwan and Ainsworth, 1992). *Task analysis* a fost utilizată, la început, pentru definirea operaţiunilor utilizator, precum şi pentru proiectarea interfeţelor utilizator şi de sistem [19], [38].

Task analysis *"... este un set de tehnici care pot fi utilizate pentru a determina care activitate a utilizatorului se doreşte a fi îndeplinită, cum se planifică îndeplinirea ei şi cum este ea îndeplinită – informaţii care plasează datele în context şi plasează utilizatorul pe primul loc în procesul de formulare de cerinţe*" [38].

Ca prim pas către integrarea inteligibilă a GVis, GIS şi KDD, este sugerat procesul de construcţie şi înlănţuire de instrumente care suportă un concept ***metodă-activitate-operaţie***. *Metodele* sunt de forma celor descrise în secţiunile anterioare (a se vedea tabelul 3 pentru o trecere în revistă) şi stabilesc "cum" se poate face explorarea în vederea atingerii scopului/scopurilor conceptual/conceptuale. O *activitate* este o frază conţinând "ce" trebuie îndeplinit prin structurarea unei organizări ierarhice sau secvenţiale de sarcini. O *operaţie* este acţiunea elementară perceptuală, motorie sau cognitivă, a cărei execuţie este necesar să fie îndeplinită de activităţi.

Tabelul 3: Trecere în revistă a taxonomiei dezvoltate în GVis, GIS şi KDD (preluare de la Monica Wachowicz, "Integrating GVis, GIS and KDD for Exploring Spatio-Temporal Data")

GVis *Pattern Identification Model MacEachren and Ganter (1990)*	**GIS** *Virtual GeoData Model, Albrecht (1996)*	**KDD** *Process Model* [14]
Feature Identification focusing, sequencing, multivariate glyphs,	**Search** interpolation, search-by-region, search-by-attribute	**Classification** symbolic methods, statistical methods
space- time cubes, small multiples, animation	**Location Analysis** buffer, corridor, overlay, Voronoi/Thiessen	**Clustering** rule-based, set functions optimisation
Feature Comparison scatterplot matrices, parallel coordinate plots,	**Terrain Analysis** slope, catchment areas	**Summarisation** data cube, attribute-oriented induction
small multiples, map overlay, multivariate colour schemes, linking brushing	**Distribution/Neighbourhood** proximity, nearest neighbour	**Predictive Modelling** neural networks, induction trees
Feature Interpretation cone tress, spider diagrams,	**Spatial Analysis** Multivariate analysis, pattern/dispersion	**Change and Deviation Detection** Bayesian change detection
spatialisation of information	**Measurements** distance, area, volume, fractal dimension	

Nivelul de implementare

La nivelul de implementare, opţiunile se fac în legătură cu algoritmii potriviţi, care susţin activităţile, şi în legătură cu mediile software/hardware specifice, în care se realizează operaţiile. În rezumat, implementarea care suportă o metodă-activitate-operaţie la nivel operaţional ar trebui să furnizeze instrumente interactive cum sunt:

- operaţii "*drag*" şi "*drop*" pentru a crea interogări de la vizualizări şi vizualizări de la interogări;
- limbaje vizuale de interogare pentru manipularea directă a datelor;
- sesiuni de explorare constând din interogări şi vizualizări care pot fi salvate independent de orice date şi pot fi reutilizate într-un set diferit de date;
- navigare printre obiecte multiple.

Scopul principal este acela de a furniza comunicaţie transparentă între mediile software, în care s-a propus ca metodele GIS, GVis şi KDD să fuzioneze cu instrumente inteligibile. Obţinerea acestei integrări de la nivelul de implementare va cere regândirea procesului de dezvoltare GIS în ideea de a se ocupa de aspecte critice cum sunt cele legate de modul în care utilizatorii de informaţie geografică vor explora seturi mari de date, căutând şabloane de date spaţiotemporale printr-un singur mediu de explorare de date. La un moment dat, s-a obţinut implementarea de bază a mişcării formatelor de date către instrumente GVis pentru a permite utilizatorilor să acceseze direct bazele de date cu informaţie geografică. Din păcate, o dată ce instrumentul GVis a remis o vizualizare tridimensională sau animaţie, este, în general, imposibil să se activeze interogări sau algoritmi pentru *data mining* de la dispozitivul de vizualizare. Prin urmare, eforturile au fost în zona de expandare şi integrare a diferitelor sisteme într-un mediu singular, care va cere standarde de date GIS deschise (open GIS data standards), legături software, funcţii/operaţii şi reţele de comunicaţii.

3.4. SPIN! Spatial Mining for Data of Public Interest

SPIN! a fost lansat şi finanţat de către Comisia Europeană în cadrul Programului Cadru 5 prin contractul IST-1999-10536 SPIN!

Conţinutul acestei secţiuni este preluat din materialul disponibil la adresa: http://www.ccg.leeds.ac.uk/spin/overview.html.

Contextul de lucru

Pătrunderea rapidă pe piaţă a tehnologiilor pentru *Data Mining* şi *Geographic Information Systems* (GIS) este dirijată de presiunea venită din partea sectorului public, agenţiilor de mediu şi din partea sectorului industrial pentru furnizarea de soluţii inovatoare pentru un spectru larg de probleme cum sunt de ex: servicii pentru asigurarea sănătăţii publice, agenţii de mediu, care evaluează impactul schimbării folosinţei terenului datorită schimbării climei, companii de geo-marketing care studiază segmentarea clienţilor/consumatorilor pe baza localizării spaţiale.

Pentru suportarea acestor tipuri de analize, majoritatea GIS-urilor au doar funcţionalitate de bază pentru analiză spaţială. Multe se limitează la analize care implică afişare statistică descriptivă cum ar fi histograme sau *pie charts*. *Data mining* care este căutare parţial automatizată, de şabloane ascunse în baze de date mari, oferă potenţiale beneficii pentru aplicarea GIS ca bază pentru luarea deciziei în organizaţiile din sectorul public şi privat.

SPIN! este unul dintre cele mai inteligibile şi mai ambiţioase proiecte care şi-a propus să pună laolaltă unele dintre cele mai interesante abordări din domeniul *data mining* şi cartografie interactivă.

Scopuri

Obiectivul principal al proiectului SPIN! este acela de a oferi posibilităţi pentru analizarea datelor georeferenţiate. În acest scop, a fost dezvoltat sistemul Spatial Data Mining, care integrează nivelul de funcţionalitate la care s-a ajuns până acum în GIS şi *Data Mining*, într-o arhitectură deschisă, extensibilă, *internet-enabled plug-in*. Domeniul *Data Mining* va progresa prin adaptarea metodelor din *Machine Learning* şi *Bayesian Statistics* la analiza spaţială. Domeniul GIS va progresa prin dezvoltarea de noi metode pentru vizualizarea informaţiei temporale şi spaţiale. Sistemul SPIN! pentru *spatial mining* va fi testat şi evaluat în cadrul unor aplicaţii dedicate analizei seismice şi vulcanologice şi pentru diseminarea bazată pe web a datelor de recensământ.

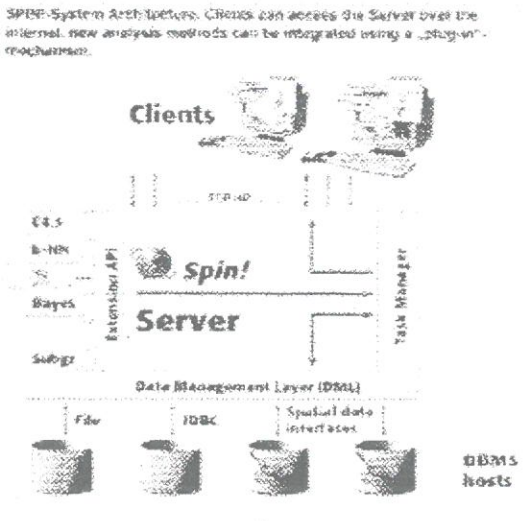

Figura 5: Arhitectura SPIN! (preluată din materialul disponibil la adresa: http://www.ccg.leeds.ac.uk/spin/overview.html)

În ultimii ani, câţiva dintre partenerii proiectului au dezvoltat componente tehnologice şi instrumente ştiinţifice cerute de nucleul unui astfel de sistem. Pe parcursul acestui proiect aceste eforturi individuale şi expertiza asociată se vor reuni la nivel european. În această idee, vor fi dezvoltate pisele care lipsesc şi va fi construită o platformă integrată GIS-Data-Mining. Partenerii din sectorul industrial vor dezvolta un model de business pentru brokering bazat pe web, cu date statistice georeferenţiate, şi vor face estimări privind impactul economic al tehnologiei.

Mai multe detalii despre cele prezentate în acest capitol se găsesc în Raportul tehnic 630/2, elaborat în mai 2002: "*Some key aspects in geospatial data mining*" [36], iar aspecte legate de tehnicile şi algoritmii folosiţi în *geospatial data mining* sunt prezentate în Raportul tehnic 630/1 elaborat în aprilie 2002: "*Techniques and algorithms in geospatial data mining*" [35].

4. Concluzii şi tendinţe

Progresele făcute, în ultimii zece ani, în domeniul teledetecţiei şi capturării de date spaţiale/geografice au sporit în mod dramatic posibilităţile de colectare zilnică de informaţie geografică la nivel de terabytes. Cu toate acestea, bunăstarea datelor geografice nu poate fi complet realizată atunci când informaţia implicită din date este dificil de separat, ceea ce duce la confruntarea specialiştilor în informaţie geografică cu o cerere urgentă de metode şi de instrumente noi, care pot să transforme în mod inteligent şi automat *datele* geografice în *informaţii* şi să sintetizeze *cunoştinţe* geografice. Trebuie să se apeleze la noi abordări în reprezentarea geografică, procesarea de interogări, analiza spaţială şi vizualizarea de date [18], [46], [62]. Specialiştii în informatică sunt confruntaţi cu aceeaşi provocare ca rezultat al revoluţiei digitale ce expediază date de ordinul terabytes de la carduri de credit tranzacţionale, examinări medicale, apeluri telefonice şi de la alte numeroase activităţi umane. Eforturile comunităţilor din domeniul inteligenţei artificiale, statistică şi baze de date, au sprijinit tehnologiile KDD în extragerea de informaţii din masivele de date pentru asigurarea suportului în luarea deciziei [2], [24], [31].

În prima parte a acestui capitol, a fost identificat potenţialul impact asupra *geographic information science* şi asupra cercetării în sens mai larg. În partea a doua, este prezentată o listă de subiecte de cercetare, oferită de Miller and Han [46].

4.1. Impactul potenţial asupra *geographic information science* şi asupra cercetării

Există cerinţe unice şi provocări în ceea ce priveşte descoperirea de cunoştinţe geografice, în cadrul geographic information science. Majoritatea bazelor de date digitale sunt, în cel mai bun caz, o reprezentare foarte simplă a cunoştinţelor geografice la nivelul geometriei elementare, constrângerilor topologice şi provenite de la măsurători. Se estimează că GIS-urile bazate pe cunoştinţe vor construi cunoştinţe geografice de nivel înalt, în cadrul bazelor de date geografice digitale, pentru analizarea fenomenelor complexe [58], [61]. Descoperirea de cunoştinţe geografice este o sursă potenţială bogată, pentru GIS bazat pe cunoştinţe şi pentru analiză spaţială inteligentă. Dezvoltarea reprezentărilor cunoştinţelor geografice descoperite, care sunt efective în GIS bazate pe cunoştinţe şi analiză spaţială, este una din abordările critice provocatoare.

Descoperirea de cunoştinţe geografice în cercetarea geografică

Informaţia geografică a fost întotdeauna un obiect de uz curent central în cercetarea geografică. Din punct de vedere istoric, cercetarea geografică a apărut în medii sărace în date. Multe dintre progresele din cercetarea geografică vor sta la baza îmbunătăţirilor tehnologiilor pentru georeferenţiere, capturare, memorare şi procesare de date geografice. Revoluţia produsă în domeniul datelor geografice digitale poate fi considerată ca fiind cea mai dramatică schimbare de atitudine în mediul dedicat cercetării geografice din istoria ştiinţei. Aceasta generează, poate, cea mai importantă "meta-chestiune" pentru cercetarea geografică a acestui secol şi anume "*care sunt problemele la care nu s-a găsit un răspuns până acum?*"

Ne aflăm încă la începuturile istoriei descoperirii de cunoştinţe geografice. În acest moment, în opinia noastră, poate fi oferită doar o listă a aplicaţiilor *geographic knowledge discovery* în *geographic information science* şi, în general, în cercetarea geografică.

- **Interpretarea hărţilor şi extragerea de informaţii**

Malebra ş.a. [43] au demonstrat utilizarea de instrumente *inductive machine learning* în cadrul mediilor GIS. Sistemul lor poate să extragă şi să interpreteze caracteristici umane şi fizice complexe din hărţi topografice pentru introducerea lor într-un GIS şi pentru analiză.

- **Extragerea de informaţie de la imaginile provenite din teledetecţie**

Creşterea rezoluţiilor spaţiale, temporale şi spectrale, furnizată de progresele făcute de tehnologiile specifice teledetecţiei, a condus la crearea de masive de baze de date de imagini. Aceste baze de date dau cercetătorilor posibilitatea de a analiza şi înţelege informaţia din aceste date. Gopal ş.a. [29] au folosit reţelele neuronale artificiale, combinate cu tehnici de vizualizare, pentru interpretarea şi înţelegerea şabloanelor extrase din imaginile provenite de la teledetecţie.

- **Carcateristicile de mediu ale hărţilor**

Multe dintre fenomenele geografice au atribute complexe, multidimensionale, care sunt dificil de rezumat şi de integrat folosind metodele analitice tradiţionale. Eklund ş.a. [13] au folosit tehnici de inductive learning şi reţele neuronale artificiale pentru a clasifica şi reprezenta pe hartă tipuri de sol. Lees ş.a. [41] au folosit metode de inducţie, bazate pe arbori de decizie, pentru reprezentarea pe hartă a tipurilor de vegetaţie, în zonele în care metodele de clasificare din teledetecţie confundă terenul cu tulburările neobişnuite (de ex. incendii).

- **Extragerea de şabloane spaţio-temporale**

Identificarea şabloanelor neobişnuite din masivele de baze de date spaţiotemporale poate fi dificilă, iar numărul de şabloane posibile poate fi foarte mare. Mesrobian ş. a. [44] au dezvoltat Open Architecture Scientific Information Sistem (OASIS) pentru interogarea, explorarea şi vizualizarea fenomenelor geofizice din baze de date mari, eterogene, distribuite. Componenta Conquest Scientific Query Processing Sistem a lui OASIS, identifică activitatea ciclonică din datele despre vreme şi climă prin extragerea de şabloane neobişnuite din presiunea aerului şi vânt în timp. În alte domenii, Openshaw şi colegii lui [52], [53] au dezvoltat tehnici exploratorii, bazate pe metode simple de interogare, pentru afişarea de clustere spaţiotemporale în datele criminalistice.

- **Interacţiune, flux şi mişcare**

Interacţiunea spaţială, fluxul şi mişcarea în spaţiul geografic pot să furnizeze noi puncte de cercetare în structura spaţială a sistemelor geografiei fizice şi umane. Structura spaţială şi interacţiunea spaţială sunt strâns legate: locaţia influenţează şabloanele de intearcţiune, în timp ce şabloanele de intearcţiune influenţează locaţia entităţilor şi a activităţilor. Din raţiuni de maleabilitate, analiza spaţială şi reţeaua analitică formulează ipoteze puternice în legătură cu influenţele asupra fluxului, intearcţiunii, mişcării şi localizării, doar prin capturare directă şi efecte de aproximare în spaţiu şi timp. Influenţele de ordin *n* pot fi îngropate în masive de baze de date de interacţiune, fluxul şi mişcarea fiind capturate prin sisteme de monitorizare în timp real, sisteme de transport inteligente şi dispozitive "*position-aware*" cum sunt telefoanele celulare şi clienţii *wireless Internet*. Marble et al. (1997) au descris metode de vizualizare pentru a explora matrici de interacţiune masive. Smyth (2001) a explorat posibilităţile pentru descoperire de cunoştinţe geografice din traiectoriile spaţiu-timp ale dispozitivelor mobile.

4.2. Aspecte critice în provocările cercetării

Există numeroase aspecte critice în descoperirea de cunoştinţe geografice şi data mining. Miller şi Han [46] oferă următoarea listă de subiecte de cercetare în acest domeniu:

- **Dezvoltarea şi suportul pentru geographic data warehouses**

Trebuie spus că nu există o adevărată *geographic data warehouse* (GDW). Proprietăţile spaţiale sunt deseori reduse la atribute spaţiale simple în cadrul *data warehouses*. Crearea de GDW integrate cere rezolvarea unor probleme de interoperabilitate spaţială şi temporală a datelor, incluzând diferenţele din semantică, sistemele de referinţă, geometrie, acurateţe şi poziţie.

- **Cele mai bune reprezentări în geographic knowledge discovery**

Tehnicile actuale de *geographic knowledge discovery* (GKD) utilizează, în general, reprezentări foarte simple ale obiectelor geografice şi ale relaţiilor spaţiale. Tehnicile pentru *geographic data mining* vor recunoaşte obiecte geografice mult mai complexe (linii şi poligoane) şi relaţii (distanţe neeuclidiene, direcţie, conectivitate şi intearcţiune în cadrul spaţiului geografic de atribute). În aceste reprezentări geografice şi relaţii trebuie să fie complet integrat timpul.

- **Geographic knowledge discovery folosind diferite tipuri de date**

Pot fi dezvoltate tehnici de *geographic knowledge discovery* GKD astfel încât să poată fi manipulate diverse tipuri de date începând cu modelele tradiţionale raster şi vector, până la imagini multimedia şi georeferenţiate precum şi tipuri de date dinamice (video streams, animaţie şi realitate virtuală).

- **Interfeţe utilizator pentru geographic knowledge discovery**

GKD necesită schimbarea de optică de la cercetătorii orientaţi tehnic către comunităţile de cercetare GIScience şi către alte comunităţi de cercetare. Această schimbare de optică necesită *interfeţe* şi instrumente care pot să ajute cercetătorii în aplicarea acestor tehnici la chestiuni specifice.

- **Verificarea conceptelor şi benchmarking**

Ca şi în alte domenii, şi în KDD şi DM trebuie să fie unele cazuri de test, definitive sau *benchmarks*, care să ilustreze puterea şi utilitatea GKD în descoperirea de cunoştinţe geografice neaşteptate. O abordare conexă este şi cea legată de *benchmarking* în vederea determinării efectelor datelor de diferite calităţi asupra cunoştinţelor geografice descoperite.

- **Construirea cunoştinţelor geografice descoperite în cadrul GIS şi analiza spaţială**

Sunt absolut necesare reprezentări ale cunoştinţelor geografice descoperite, care sunt potrivite pentru GIS şi analiză spaţială. Aceasta poate să includă interfeţe GIS bazate pe *on line analytical processing* (OLAP) şi instrumente inteligente pentru ghidonarea analizei spaţiale.

În cartea "Advances in Knowledge Discovery and data Mining", publicată de MIT Press, Usama M. Fayyad şi editorii fac următoarea remarcă: "...din combinarea celor doi termeni "data mining" şi "data warehousing",

este de aşteptat să se construiască un pod între comunităţile de specialişti din statistică, baze de date şi machine learning, care să atragă foarte mult intersul dezvoltatorilor de sisteme informatice".

Bazată pe studiul literaturii de specialitate şi pe experienţa proprie, predicţia formulată de specialişti este aceea că modul de familiarizare cu termenii "data mining" şi "data warehousing" va fi similar cu ce s-a întâmplat cu termenul "GIS" în ultimii cinci ani.

Menţiune

Acest articol este un rezumat al Raportului Tehnic, elaborat de dr. Angela Ioniţă (Institutul Naţional pentru Cercetare-Dezvoltare în Informatică – ICI Bucureşti) în mai 2002, în cadrul proiectului INTAS nr. 397: **DATA MINING TECHNOLOGIES AND IMAGE PROCESSING: THEORY AND APPLICATIONS** coordonat de Lappeenranta University of Technology din Finlanda.

Bibliografie

1. **ANSELIN, L.**: Local Indicators of Spatial Association. LISA. În:Geographical Analysis, 27, 1995.
2. **BHANDARI, E. E. COLET, J. PARKER, Z. PINES, R. PRATAP, R. PRATAP, K. RAMANUJAM**: Advanced Scout: Data Mining and Knowledge Discovery in NBA Data. În: Data Mining and Knowledge Discovery, 1, 1997.
3. **BRACHMAN, R.J., T. ANAND**: The Process of Knowledge Discovery in Databases in Advances in Knowledge Discovery and data Mining, U.M. Fayyad et al. (eds.) AAAI Press/ The MIT Press, 1996.
4. **BRADSIL, P. B. K. KONOLIGE** (Eds.): Meta-Learning, Meta-Reasoning and Logics, Kluwer Academic Press., Boston, MA, USA, 1990.
5. **BRUNSDON, C., A. S. FOTHERINGHAM, M. E. CHARLTON**: Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. În: Geographical Analysis, 28, 1996.
6. **BUTTENFIELD, B.P.**: Looking Forward: Geographic Information Services and Libraries in the Future. În: Cartography and GIS, 25(3), 1998.
7. **BUTTENFIELD, B.P., M. GAHEGAN, H. MILLER, M. YUAN**: Geospatial Data Mining and Knowledge Discovery.
8. **CÂMARA, A. S. J. RAPER** (Eds.): Spatial Multimedia and Virtual Reality, London: Taylor and Francis, 1999.
9. **CANTU-PAZ, ERIK** and **CHANDRIKA KAMATH**: On the Use of Evolutionary Algorithms in Data Mining.
10. **CHEESEMAN, P. J. STUTZ**: Bayesian Classification: Theory and Results. În: Fayyad, U., Piatetsky-Shapiro, G, Smyth, P. and Uthurusamy, R. (Eds.) Advances in Knowledge Discovery and Data Mining, Cambridge, MA: AAAI/MIT Press, 1996.
11. **CHEN, M., J. HAN, P.S. YU**: Data Mining – An Overview from a Database Perspective: Focussing Techniques for Efficient Class Identification. În: Proc. of Int. Symposium on large databases (SSD'95), Maine, 1996.
12. **EGENHOFER, M. J. J.R. HERRING**: Categorizing Binary Topological Relations Between Regions, Lines and Points in Geographic Databases. În: M. Egenhofer, D. M. Mark and J. R. Herring (Eds.), The 9-intersection: Formalism and its Use for Natural-language Spatial Predicates, Santa Barbara, CA: National Center for Geographic Information and Analysis Technical Report 94-1), 1994.
13. **EKLUND, P. W., S. D. KIRKBY, S. D., A. SALIM**: Data mining and soil salinity analysis. În: International Journal of Geographical Information Science, 12, 1998.
14. **FAYYAD, U., G. PIATETSKY-SHAPIRO, P. SMYTH**: From Data Mining to Knowledge Discovery in Databases. În: AI Magazine, Fall, 1996.
15. **FAYYAD, U.**: Editorial. Data Mining and Knowledge Discovery, 1997.
16. **FLEWELLING, D. M., M. J. EGENHOFER**: Using Digital Spatial Archives Effectively. În: The Int. Journal of Geographical Information Science, 1999.
17. **FOTHERINGHAM, A. S., M. CHARLTON, C. BRUNSDON**: Two Techniques for Exploring Non-Stationarity in Geographical Data. În: Geographical Systems, 1997.

18. **GAHEGAN, M.**: On the Application of Inductive Machine Learning Tools to Geographical Analysis. În: Geographical Analysis, 2000.

19. **GAHEGAN, M** Gahegan, M. and O'Brien, D. (1997). *A Strategy and Architecture for the Visualisation of Complex Geographical Datasets.* International Journal of Pattern Recognition and Artificial Intelligence, 11(2);

20. **GAHEGAN, M., M. TAKATSUKA, M. WHEELER, F. HARDISTY**: GeoVISTA Studio: A Geocomputational Workbench. În: Proc. 4th Annual Conference on GeoComputation, UK, August 2000. URL: http://www.ashville.demon.co.uk/gc2000/.

21. **GAHEGAN, M.**: National Academies White Paper. În: Intersection of Geospatial Information and Information Technology. September, 2001.

22. **GAHEGAN, M.**: Data Mining and Knowledge Discovery in the Geographical Domain.

23. **GAHEGAN, M., M. WACHOWICZ, M. HARROWER, T. M. RHYNE**: The Integration of Geographic Visualization with Knowledge Discovery in Databases and Geocomputation. În: Cartography and Geographic Information Sistems, special issue on the ICA research agenda, 2001.

24. **GARDNER, C.**: IBM Data Mining Technology, Stamford, Connecticut: IBM Cooperation, 1996.

25. **GETIS, A., J.K. ORD**: The Analysis of Spatial Association by Use of Distance Statistics. În: Geographical Analysis; 1992.

26. **GETIS, A., J.K. ORD**: Local spatial statistics: An overview. În: P. Longley and M. Batty (Eds.) Spatial Analysis: Modelling in a GIS Environment, Cambridge, UK: GeoInformation International, 1996.

27. **GOODCHILD, M. F.:** Geographical Data Modeling. În: Computers and Geosciences, Vol. 18, No. 4, 1992.

28. **GOODCHILD, M. F., B.P. BUTTENFIELD, P. ADLER, A. KRYGIEL, H. ONSRUD, R. KAHN**: Distributed Geolibraries. National Research Council Monograph., Washington, D.C.: National Academy Press, 1999.

29. **GOPAL, S., W. LIU, C. WOODCOCK**: Visualization Based on Fuzzy ARTMAP Neural Network for Mining Remotely Sensed Data. În H. J. Miller and J. Han (Eds.) Geographic Data Mining and Knowledge Discovery, London: Taylor and Francis, 2001 (in press).

30. **GUNOPULOS, D.**: Data mining techniques for geospatial applications.

31. **HEDBERG, S. R.**: Search for the Mother Lode: Tales of the First Data Miners. În: IEEE Expert, 11(5), 1996.

32. **HEALEY, R.**: Database Management Sistems. În: Maguire, D., Goodchild, M.F., and Rhind, D., (Eds.), Geographic Information Sistems: Principles and Applications, London: Longman, 1991.

33. **HINKE, T., J. S. RUSHING, S. KANSAL, GRAVES, AND H. RANGANATH**: For Scientific Data Discovery: Why Can't the Archive be More Like the Web. În: Proc. Of the 9th Int. Conf. on Scientific Database Management, Olympia, WA, Aug. 11-13, 1997.

34. **HORNSBY, K., M. J.EGENHOFER**: Identity-based Change: A Foundation for Spatio-Temporal Knowledge Representation. În: International Journal of Geographical Information Science, 14, 2000.

35. **IONIȚĂ, A.**: Techniques and Algorithms in Geospatial Data Mining. Technical Report, TR ICI 630/1, April 2002.

36. **IONIȚĂ, A.**: Some Key Aspects in Geospatial Data Mining. Technical report, TR ICI 630/2, May, 2002.

37. **JACKSON, J.E.**: An User's Guide to Principial Components, New York, NY:John Wiley, 1991.

38. **KNAPP, L.:** A Task Analysis Approach to the Visualisation of Geographic Data. În: Cognitive Aspects of Human-Computer Interaction for Geographic Information Sistems, T.L. Nyerges et al. (Eds.), Kluwer Academic Publishers, 1995.

39. **KOPERSKI, K., J. HAN, J.:** Discovery of Spatial Association Rules in Geographic Information Databases. În: Proc. of the 4th International Symposium on Large Spatial Databases, SSD95, Maine, 1995.

40. **KOPERSKI, K., J. HAN, J. ADHIKARY:** Mining Knowledge in Geographic Data. În: Comm. ACM, 1999. Available at URL: http://db.cs.sfu.ca/sections/publication/kdd/kdd.html

41. **LEES, B. G. K. RITMAN**: Decision-Tree and Rule-Induction Approach to Integration of Remotely Sensed and GIS Data in Mapping Vegetation in Disturbed or Hilly Environments. În: Environmental Management, 15, 1991.

42. **MACEACHREN, A.M., M. J. KRAAK**: Exploratory Cartographic Visualization: Advancing the Agenda. În: Computers and Geoscience (4); 1997.

43. **MALEBRA, D., F. ESPOSITO, A. LANZA, F. LISI**: Machine Learning for Information Extraction from Topographic Maps. În: H. J. Miller and J. Han (Eds.), Geographic Data Mining and Knowledge Discovery, London: Taylor and Francis, 2001 (in press).

44. **MESROBIAN, E, R. MUNTZ, E. SHEK, S. NITTEL, M. LA ROUCHE, M. KRIGUER, C. MECHOSO, J. FARRARA, P. STOLORZ, H. NAKAMURA:** Mining Geophysical Data for Knowledge. În: IEEE Expert, 1996.

45. **MILLER, H. J.:** Geographic Representation in Spatial Analysis. În: Journal of Geographical Sistems, 2, 2000.

46. **MILLER, H. J., J. HAN:** Discovering Geographic Knowledge in Data Rich Environments: A report on a specialist meeting, SIGKDD Explorations: Newsletter of the, Association for Computing Machinery, Special Interest Group on Knowledge Discovery and Data Mining, 1(2), 2000, available at http://www.acm.org/sigs/sigkdd/explorations.

47. **MILLER, H. J., J. HAN (Eds):** Geographic Data Mining and Knowledge Discovery, London: Taylor & Francis, 2001.

48. **MILLER, H. J., H. JIAWEI:** Geographic Data Mining and Knowledge Discovery: An Overview. În: H. J. Miller and J. Han (Eds.) Geographic Data Mining and Knowledge Discovery, London: Taylor and Francis, 2001 (in press).

49. *** * *:** NPR National Public Radio Morning Edition. Report on Emergence of Commercial Internet Search Engines such as Yahoo and Other .dot-coms., 3 April, 1998.

50. *** * *:** NRC Data Foundation for the National Spatial Data Infrastructure. National Research Council Mapping Science Committee, Sugarbaker, L.A., Chair, Washington, D.C.: National Academy Press, 1995.

51. **OKABE, A, H. J. MILLER:** Exact Computational Methods for Calculating Distances Between Objects in a Cartographic Database. În: Cartography and Geographic Information Sistems, 1996.

52. **OPENSHAW, S.:** The Modifiable Areal Unit Problem. CATMOG 38, Norwich: Geo Abstracts, 1984.

53. **OPENSHAW, S.:**, Two Exploratory Space-time-attribute Pattern Analysers Relevant to GIS. În: A. S. Fotheringham and P. A. Rogerson (Eds.), Spatial Analysis and GIS, London: Taylor and Francis, 1994.

54. **OPENSHAW, S.:** Geographical Data Mining: Key Design Issues, 1999, available at the address: http://www.geovista.psu.edu/sites/geocomp99/Gc99/051/gc_051.htm

55. **PEUQUET, D. J., C. X. ZHANG**: An Algorithm to Determine the Directional Relationship Between Arbitrarily-shaped polygons in the Plane. În: Pattern Recognition; 1987.

56. **PEIRCE, C. S.:** Deduction, Induction and Hypothesis. În: Popular Science Monthly, 13, 1878.

57. **RODDICK , J. F., B. LEES**: Paradigms for Spatial and Spatio-temporal Data Mining. În: H. J. Miller and J. Han (Eds.)., Geographic Data Mining and Knowledge Discovery, London: Taylor and Francis, 2001 (in press).

58. **SRINIVASAN, A., J. A. RICHARDS**: Analysis of GIS Spatial Data Using Knowledge-based Methods. În: International Journal of Geographical Information Sistems, 7, 1993.

59. **WACHOWICZ, M.**: Integrating Gvis, GIS and KDD for Exploring Spatio-Temporal Data.

60. **WATTERSON, K.:** A Data Miner's Tools. În: BYTE, October 1995.

61. **YUAN, M.**: Use of Knowledge Acquisition to Build Wildfire Representation in Geographic Information Systems. În: International Journal of Geographical Information Sistems, 11, 1997.

62. **YUAN, M.:** Representing Spatiotemporal Processes to Support Knowledge Discovery in GIS databases. În: T. K. Poiker and N. Chrisman (Eds.), Proc. of the 8th International Symposium on Spatial Data Handling Spatial Data Handling; 1998.

63. **YUAN, M., B. BUTTENFIELD, M. GAHEGAN, H. MILLER:** Geospatial Data Mining and Knowledge Discovery. A UCGIS White Paper on Emergent Research Themes, 2001. URL: http://www.ucgis.org/emerging/.

64. **ZAÏANE, O. R., J. HAN, Z-N. LI, J. HOU**: Mining Multimedia Data. Proceedings, CASCON'98: Meeting of Minds, Toronto, Canada, November 1998; available at: http://db.cs.sfu.ca/sections/publication/smmdb/smmdb.html

65. http://db.cs.sfu.ca/DBMiner

66. http://db.cs.sfu.ca/DBMiner