ADATBÁNYÁSZAT ÉS STATISZTIKA

DR. SRAMÓ ANDRÁS

A 70-es évek közepétől napjainkig eltelt időszak drámai növekedést hozott az elektronikus adattárolásban. Az automatizált, illetve diverzifikált adatrögzítési technológiák elterjedésével robbanásszerűvé vált ez a növekedés a 90-es években. Információtechnológiai szempontból említésre méltó például a vonalkódok alkalmazása, a pénztárgépek összekapcsolása egy központi számítógéppel, bankautomaták elterjedése stb. Becslések szerint a világban hasznosított információ húszhavonta megkétszereződik, az alkalmazott adatbázisok száma és mérete pedig még ennél is gyorsabban növekszik.

Ahogy azonos árszínvonalon egyre gyorsabb számítógépek, valamint egyre nagyobb tárolókapacitás érhető el, az adattárolás költsége exponenciális mértékben csökken, ennek eredményeképpen az adatok egyre olcsóbbá válnak. A szervezeti adatbázisok méretének növekedésével, a különböző funkcionális (számviteli, pénzügyi, marketing stb.) adatbázisok összekapcsolásával egy új fogalom született, az adattárház (data warehouse). Az adattárház lényege a különböző belső és külső forrásokból származó nyers adatok integrálása egy olyan rendszerben, amely egységes erőforrásként használható a szervezet végfelhasználói – legtöbbször a szervezetek vezetői – számára. Az adattárház végfelhasználói ad hoc lekérdezéseket és on line elemzéseket hajthatnak végre a nem ritkán gigabájtos méretet elérő adathalmazon. Az adattárházat napjainkban már a döntéstámogatás alapjának tekintik [10].

A nagymértékű adatkoncentráció ráirányította a figyelmet arra a kérdésre, hogy mit lehet még – az eredeti feldolgozási célokon túl – tenni ezzel az értékes adattömeggel. Mivel az információ az üzleti tevékenység erőforrásaként kezelendő, a döntéshozatalt a szervezeti adatbázisokból elérhető információkra kell felépíteni. A hagyományos on line tranzakciófeldolgozó-rendszerek képesek a szervezeti adatbázisok gyors, biztonságos és hatékony feltöltésére, de nem a legjobbak az adatelemzésben: előre megfogalmazott kérdésekre válaszolnak, vagy interaktív módon biztosítják, hogy a döntéshozó tegyen fel kérdéseket a rendszerben tárolt adatokra vonatkozóan.

Az adatelemzés jelentőségét az adja meg, hogy ha segítségével az explicit módon tárolt adatok mögé nézünk, új ismereteket szerezhetünk. Amennyiben az adatelemzés új ismeretek előállítását emeli ki, akkor ismeretkeresésről (Knowledge Discovery in Databases – KDD) vagy egy népszerűbb kifejezéssel élve, adatbányászatról (Data Mining) beszélhetünk. A szervezeti adatbázisban való ismeretkeresésnek nyilvánvaló előnyei vannak a szervezet számára.

Az adatbányászat ma már messze túlmutat az adatelemzés hagyományos formáin. Az adatbányászat, illetve az adatbázisban való ismeretkeresés néhány népszerű definícióját az alábbiakban adjuk meg:

„Az adatbányászat implicit, előzőleg ismeretlen és potenciálisan hasznos információ adatokból történő nem triviális kivonását jelenti. Az adatbányászat számos eltérő alapon nyugvó eljárást tartalmaz: osztályozás, adatösszegzés, osztályozó szabályok tanulása, függési hálók keresése, változáselemzés és anomáliakeresés stb." [3]

„Az adatbányászat olyan kapcsolatok és globális minták nagy adatbázisokban történő keresését jelenti, amelyek el vannak rejtve az adatok nagy tömege mögött, mint például páciensek adatai és az orvosi diagnózisok között fennálló kapcsolat. Ezek a kapcsolatok értékes ismereteket jelenthetnek az adatbázisról és az adatbázisok objektumairól, ha pedig az adatbázis hiteles tükörképe a valóságnak, akkor a valós világról." [7]

Az adatbányászat adatelemzéssel és különböző szoftvertechnikák alkalmazásával foglalkozik abból a célból, hogy a rendelkezésre álló adathalmazokban mintákat és szabályosságokat keressen. További cél, hogy ez az ismeretkeresés lehetőség szerint automatizált módon történjék. Így az adatbázisokból ismeretbázis építhető fel [5], [6].

„Az adatbányászat az adatokban rejlő minták, kapcsolatok, változások, anomáliák és statisztikailag szignifikáns struktúrák és események felfedezésével foglalkozik. A hagyományos adatelemzés feltevés alapú abban az értelemben, hogy egy hipotézist fogalmaz meg és annak helytállóságát igazolja vagy elveti. Az adatbányászat ezzel ellentétben felfedezés alapú annak megfelelően, hogy a mintákat automatikusan vonja ki az adatokból." [4]

Az adatbányászat egy nagyobb iteratív folyamat része, amelyet ismeretkeresésnek nevezünk. Az ismeretkeresés a nyers adatokból indul ki és az új ismeretek megfogalmazásával fejeződik be. A folyamat fázisai a következők.

– *A probléma definiálása*. Először az ismeretkeresés céljait azonosítjuk, és igazoljuk, hogy a célok elérésével felfedezett új ismeretek hasznosíthatók. Ebben a fázisban történik az adatok egy részhalmazának meghatározása úgy, hogy valamilyen kritériumnak megfelelően az adatokat kiválasztjuk vagy szegmentáljuk.

– *Adatok gyűjtése, tisztítása és előkészítése.* Az adatok különböző belső és külső forrásokból gyűjthetők össze. Az adatbányászat végrehajtásához fel kell oldani az adatok reprezentációjában és kódolásában jelentkező ellentmondásokat. Egységes adattáblázatokat hozunk létre, amelyekből ki kell szűrni a szokatlan, az egymásnak ellentmondó és a hiányzó adatokat. Új, származtatott adatok meghatározására is sor kerülhet, amelyeket a meglévő adatokból számolunk ki. Ez a leginkább munkaigényes fázis, legtöbbször a teljes feladat 70 százalékát jelenti. Ha viszont az adatok adattárházban találhatók, lényegesen kevesebb erőfeszítésre van szükség a megfelelő adathalmaz előállítására.

– *Modellalkotás*. Ebben a fázisban történik az adatbányászatra használt modell és eszköz kiválasztása, az eszköz által igényelt transzformációk végrehajtása. A modell ellenőrzésére (neurális háló alkalmazása esetén gyakorlatoztatására) mintákat kell generálni, és ezeknek a mintáknak a használatával teszteljük az eszközt és a modellt.

– *A modellek igazolása*. A modellt egy olyan független adathalmazzal kell tesztelni, amelyet nem használtunk a modellalkotásban. A tesztelés a modell pontosságára, érzékenységére és használhatóságára irányul.

– *A modell telepítése*. Egy előrejelző modell esetén a modellt új esetekre alkalmazzuk, és az előrejelzések alapján megváltoztatjuk a szervezet működését, az üzletmenetet. A modell telepítése szükségessé teheti számítógéprendszerek telepítését, amelyek valós idejű előrejelzéseket generálnak úgy, hogy a döntéshozó alkalmazkodni tudjon döntéseivel az előre jelzett jelenségekre.

– *Ellenőrzés, a modell figyelése*. Bármit is modellezünk, az biztosan meg fog változni az időben. Ezért szükség van a modell folyamatos figyelésére és lehetséges módosítására, ha egyáltalán a modell alkalmazhatósága nem kérdőjeleződik meg.

Az ismeretkeresés itt bemutatott folyamata iteratív. Bármely fázisban felvetődhet, hogy az adatok egy része használhatatlan, vagy további adattisztításra van szükség.

AZ ADATBÁNYÁSZAT ELMÉLETI HÁTTERE

Az adatbányászat tipikusan interdiszciplináris tudomány, az adatbázis-elmélet mellett itt most három hangsúlyos területet említünk meg: az induktív tanulást, a gépi tanulást és a statisztikát. Az adatbányászattal foglalkozó kutatásokat és publikációkat tekintve azt tapasztalhatjuk, hogy ez a három terület nem egyforma súllyal van jelen a kutatásokban és a kifejlesztett adatbányász-szoftverekben. *K. M. Decker* és *S. Focardi* egy 1995-ös kutatási jelentésükben csak a gépi tanulást említik mint adatbányász-technológiát, és ennek alapján tesznek javaslatot az adatbányász-módszerek egységesítésére [2].

Az indukció az adatokból információra történő következtetést jelenti, és az *induktív tanulás* az a modellépítő folyamat, amelynek során a környezetet – azaz az adatbázist – elemezzük mintakeresés céljából. A hasonló objektumokat osztályokba soroljuk, és amikor csak lehetséges, szabályokat fogalmazunk meg a közvetlenül nem megragadható objektumok osztályainak előrejelzésére. Az osztályozásnak ez a folyamata úgy azonosítja az osztályokat, hogy minden osztály az értékek egyetlen mintájával rendelkezik, és ez a minta alkotja az osztály leírását. A környezet dinamikus volta miatt a modellnek adaptívnak kell lennie, azaz képesnek kell lennie a tanulásra.

Az induktív tanulás két fő stratégiáját különböztethetjük meg.

a) *Felügyelt tanulás*: a tanulás példák alapján történik, ahol a tanító úgy segíti a rendszert egy modell kialakításában, hogy osztályokat definiál és példákat ad minden osztályra. A rendszernek az a feladata, hogy az osztályoknak egy olyan leírását találja meg, amely a példák közös tulajdonságait ragadja meg. Ha egy ilyen leírás elkészült, akkor a leírás és az osztály együtt egy osztályozó szabályt alkot, amellyel korábban nem besorolható objektumok osztálya előre jelezhető. A módszer hasonló, mint a diszkriminancia-analízis a statisztikában.

b) *Nem felügyelt tanulás*: a tanulás megfigyelésből és felfedezésből történik. Az adatbányász-rendszer csak objektumokat kap osztályok nélkül, és a rendszer feladata a példák megfigyelése és a minták – azaz az osztályleírások – azonosítása. A rendszer osztályleírásoknak egy halmazát állítja elő, a környezetben felfedezett mindegyik osztály számára egyet. A módszer hasonló, mint a klaszterezés a statisztikában.

Mindezek alapján az indukció minták kivonását jelenti. Az induktív tanulási módszerek által előállított modellek minősége lehetővé teszi, hogy a modellt jövőbeli szituációk következményeinek előrejelzésére használjuk, azaz nemcsak a megfigyelt állapotokra, hanem az előre nem látható állapotokra nézve is tartalmaz megállapításokat. A módszer problémája, hogy a legtöbb környezetnek folyamatosan változó, különböző állapotai vannak, és így nem mindig lehetséges a modell igazolása minden valószínűsíthető állapotra.

A *gépi tanulás* a tanulási folyamat automatizálását jelenti, és a tanulás egyenértékű a környezeti állapotok és átmenetek megfigyelésein alapuló szabályok létrehozásával. Ez egy igen kiterjedt tudományterület, ahová nemcsak a példákból való tanulás, hanem a megerősítésen alapuló tanulás, a tanárral való tanulás is tartozik. A gépi tanulás a korábban megismert példákat és következményeiket vizsgálja, és azt tanítja meg, hogyan lehet ezeket reprodukálni, és hogyan lehet új esetekre nézve általánosításokat megfogalmazni.

Általában a gépi tanuló rendszerek nem eseti megfigyeléseket használnak a környezetükre vonatkozóan, hanem egy teljes és véges halmazt, amelyet tanulóhalmaznak nevezünk. Ez a halmaz példákat tartalmaz, azaz olyan megfigyeléseket, amelyeket a gép szá-

mára olvashatóvá kódoltak. A tanulóhalmaz véges, ebből következik, hogy nem minden következtetés tanulható meg pontosan.

Az adatbázisokban való ismeretkeresés – azaz az adatbányászat – és a gépi tanulás ugyanolyan algoritmusokat használ a példákból való tanulásra, és hasonló problémákkal foglalkozik, mégis különbséget kell tennünk közöttük.

– Az adatbányászat értelmezhető ismeretek keresésével foglalkozik, miközben a gépi tanulás célja a teljesítmény javítása. Ennek megfelelően egy neurális háló optimális beállítása része a gépi tanulás gyakorlatának, de nem része az adatbányászatnak. Arra azonban vannak kísérletek, hogy a neurális hálókat adatbányászat céljaira is alkalmazzák.

– Az adatbányászat nagyon nagy, a valós világhoz ezer szállal kötődő adatbázisokat használ, miközben a gépi tanulás legtöbbször kisebb adathalmazokkal dolgozik. Így az adatbányászat számára a hatékonysággal kapcsolatos kérdések kiemelkedő jelentőségűek.

A *statisztika* szilárd elméleti alapokkal rendelkező tudomány, de a statisztika eredményeit bonyolult értelmezni, és a legtöbb (nem statisztikus) szakember támogatást igényel abban, hogyan elemezze a rendelkezésére álló adatokat. Az adatbányászat művelése során a feladat olyan szoftverek biztosítása, amelyek egy szakértő ismereteit és fejlett számítógépes elemző technikákat együttesen bocsátanak a felhasználó rendelkezésére.

A ma elérhető statisztikai elemző rendszerek – mint például a SAS vagy az SPSS – segítségével szokatlan minták fedezhetők fel, és különböző statisztikai modellek alkalmazásával ezek a minták megmagyarázhatók. A statisztika és az adatbányászat kapcsolatát leginkább ott ragadhatjuk meg, hogy az adatbányászat során közvetlenebb és automatizált elemzésekre van szükség, azaz a feladatot nem a (statisztikus) elemző fogalmazza meg, hanem az adatbányászást végrehajtó rendszer.

Nehéz meghúzni a határvonalat az adatbányászaton belül a statisztika és a másik két itt említett tudományterület – azaz az induktív tanulás és a gépi tanulás – között, mivel vannak olyan tanulási módszerek, amelyek tisztán statisztikai módszereket alkalmaznak.

Egyes szakemberek szerint – lásd például [9] – az adatbányászat nem más, mint a statisztika kibővítése némi mesterséges intelligenciával és gépi tanulással. Mivel a statisztika nem ad kész üzleti megoldásokat, a gazdaság szereplőinek legtöbbször gondot jelent az adatbázisokban tárolt adatokon alapuló statisztikai elemzések értékelése, bevonása a döntéshozatalba. Az adatbányász-szoftverekben megtestesülő technológia azonban az adatelemzést érthetőbbé teszi, illetve automatizálja.

A statisztikán belül is izgalmas az oksági kapcsolatok vizsgálata. Mivel ezek az oksági kapcsolatok a valós világban véletlenszerűek, nyilvánvaló a statisztikai eljárások alkalmazása az összefüggés- és kapcsolatkeresésben. Az adatbányászat igényei azonban túlmutatnak a jól bejáratott korreláció- és regresszió-számításon, például a következő kérdés vizsgálatával: ha egy automatizált összefüggés-kereső módszer nem talált kapcsolatot a változóink között, állíthatjuk-e, hogy nincs a változók között összefüggés a megfigyelések egyetlen részhalmazán sem. Hasonlóan izgalmas kérdés az idősorok vizsgálata, amely napjainkig szinte kizárólag statisztikai eszközökkel történt. A speciális, időt megragadó (temporal) adatbázisokban az adatbányászat új kérdéseket vet fel, úgymint

– az adatok időbeni változásának jellemzése,
– osztályozás és csoportosítás időben változó adatok alapján,
– tipikus trendek keresése, a tipikus trendtől való eltérés vizsgálata,
– statikus és időben változó adatok szétválasztása.

MODELLEK AZ ADATBÁNYÁSZATBAN

Az adatbányászat által alkalmazott módszereket, eljárásokat szinte lehetetlen felsorolni, bár erre történnek kísérletek. Módszertanilag áttekinthetőbb rendszert kapunk, ha az adatbányászat során alkalmazott modelleket akarjuk megnevezni: az adatbányászatot bemutató tanulmányok [1], [11] is ezt teszik.

Az alkalmazott modelleknek két fajtáját különböztethetjük meg. Az *előrejelző modellek* olyan bemenő adatokat használnak a modellek kialakításában, amelyek következményei ismertek, majd ezeket a modelleket olyan adatokra alkalmazzák, amelyek következményei ismeretlenek. A *leíró modellek* mintákat, kapcsolatokat írnak le a vizsgálatba vont adatokban, amely mintákat különböző döntésekben lehet felhasználni. Az alapvető különbség a két szemlélet között, hogy az előrejelző modellek explicit előrejelzéseket fogalmaznak meg, mialatt a leíró modellek előrejelző modellek kialakításában jelentenek segítséget. Az így definiált két kategória nem különül el élesen egymástól, mivel a legtöbb előrejelző modell egyben leíró is.

A megoldandó feladat szempontjából a modelleknek hat típusa határozható meg: az osztályozás, a regresszió-számítás, az idősorelemzés, a klaszterezés, a kapcsolatelemzés és a sorozatkeresés. Az osztályozás, a regresszió-számítás és az idősorelemzés tipikusan előrejelző modellek, míg a klaszterezés, a kapcsolatelemzés és a sorozatkeresés inkább a leíró modellek kategóriájába tartozik.

A modellek alkalmazására különböző algoritmusok szolgálnak. Egy konkrét algoritmus több modelltípus létrehozását is lehetővé teszi: például a neurális hálókat ma már regresszió-számításra is használják.

A továbbiakban röviden összefoglaljuk az egyes modelltípusok alapvető feladatait.

Az *osztályozás* esetleírások azon jellemzőit azonosítja, amelyek meghatározzák, hogy az eset melyik előre definiált kategóriába (csoportba) tartozik. A modell használható az adathalmaz megismerésére, de előrejelzésre is. A modellnek stabil statisztikai háttere van (például diszkriminancia-analízis), az alapvető kérdés azonban az, hogyan származtatjuk a csoportokat. Az egyik lehetőség történeti adatbázisok alkalmazása (például különböző szolgáltatásokat igénylő ügyfelek jellemzőit keressük), vagy szakértőt alkalmazunk az adatbázis egy részének osztályozására, és ennek alapján dolgozzuk ki a modellt a teljes adatbázisra. Harmadik lehetőség a kísérlet: egy címlistából kiválasztott mintában szereplőknek levelet küldünk, és a válaszolók jellemzői alapján építjük fel a modellt.

A *regresszió-számítás* szintén régi statisztikai eljárás, különösen ha a lineáris modellt tekintjük. A valóságban a probléma ott jelentkezik, hogy a feltárni kívánt összefüggések nem lineárisak, illetve nem tudjuk, hogy milyen jellegűek. Ilyen esetekben az adatbányászat épp az összefüggés természetére vonatkozik, amit új módszerekkel támogatnak meg: ilyen például az osztályozó és regressziós fák (Classification and Regression Trees – CART) és a neurális háló.

Az *idősorelemzés* klasszikus előrejelzési feladatai azokkal az új problémákkal bővülnek, amelyeket korábban az időadatbázisokkal kapcsolatban említettünk. További érdeklődésre számot tartó problémakör az idő szerkezetének, az időperiódusok hierarchiájának vizsgálata.

A *klaszterezés* az adatbázisban tárolt esetek csoportosítását végzi oly módon, hogy a csoportok nagyon különbözzenek egymástól, a csoportba tartozó esetek pedig nagyon

hasonlítsanak egymáshoz. A klaszterezésről köztudott, hogy nagyon szubjektív módszer az alkalmazott távolságfüggvény miatt. Így nagyon könnyen előfordulhat, hogy két adatbányász különböző távolságfüggvények használatával eltérő eredményre jut. Jogosan vetődik fel a kérdés, melyik klaszterezés a helyes. Az egyik lehetséges megoldás a problématerület szakértőjének bevonása a klaszterezésbe, aki egyrészt segíthet a megfelelő távolságfüggvény kiválasztásában, illetve „ráismer" a klaszterek használhatóságára. A korszerű adatbányász-szoftverek viszont már nem csak egyetlen klaszterező algoritmust tartalmaznak. A megfelelő modell kiválasztásában célszerű a különböző algoritmusok (például neurális háló, döntési fa vagy hagyományos klaszterező eljárás) eredményeinek összehasonlítása.

A *kapcsolatelemzés* (használatos még az „asszociáció" kifejezés is) olyan adattételeket keres, amelyek egyszerre vannak jelen egy eseményben vagy az adatbázis egy rekordjában. Például, ha „A" jelen van egy eseményben, akkor *x* százalék a valószínűsége, hogy „B" is jelen van.

A *sorozatkeresés* hasonló a kapcsolatelemzéshez azzal a különbséggel, hogy az adattételek különböző időkből származnak. Ennek megfelelően a legtöbb adatbányászszoftver együtt kezeli a kapcsolatokat és a sorozatokat, a sorozatokat olyan kapcsolatoknak tekintve, amelyek az idővel vannak összekötve.

AZ ADATBÁNYÁSZAT ÉS AZ ADATOK MINŐSÉGE

Az adatminőség kérdése különösen fontos több száz gigabájtnyi méretű adatbázisokban. Az ismeretkeresésbe bevont adathalmazok minősége nyilvánvalóan meghatározza a felfedezett összefüggések megbízhatóságát. A legtöbb adatbányász-technika képes arra, hogy számokban is kifejezze a levont következtetések jelentőségét. Az így meghatározott értékeknek a felhasznált adatforrások a priori minőségét is jellemezniük kell. Adatbázis-alkalmazásokban a lekérdezésekre adott válaszok minősége alapvető fontosságú az adatbázis-felhasználók számára, míg döntéstámogató rendszerekben a támogatásba bevont adatok minősége a döntési folyamat kulcsfontosságú összetevője. Általánosságban azonban azt mondhatjuk, hogy az adatbázis-rendszerek nem felelősek a bennük tárolt adatok minőségéért, az általuk megfogalmazott válaszok értelmezését a felhasználókra hagyják.

Olyan alkalmazásokban, amelyek az információkat több, egymást átfedő forrásból állítják elő, a minőségbecslésnek fel kell oldania a kereszthivatkozásokból származó inkonzisztenciát. Ha az információt árucikknek tekintjük, akkor egy információs termék minősége a legfontosabb paraméter az információ hasznosságának és árának a meghatározásában. Például egy termék potenciális fogyasztóinak címeiből álló lista értékének arányosnak kell lennie a lista pontosságával és teljességével.

Az adatminőséggel kapcsolatos kutatási és fejlesztési kérdéseket a következőképpen csoportosíthatjuk.

– Az adatminőséget meghatározó szabvány létrehozása, amelynek alapján az információs termékek a minőségük szerint értékelhetők. Egy ilyen szabványnak nemcsak a teljes körű adatminőség meghatározására kell kiterjednie, hanem azokra a helyzetekre is, amikor az adatminőség nem homogén, azaz az adathalmazban különböző minőségű részhalmazok találhatók.

– Hatékony és megbízható eljárások fejlesztése az adatminőség meghatározására. Az eljárások között lehetnek olyanok, amelyek „laboratóriumi" körülmények között működtethetők, azaz a minőségbecslés felhaszná-

lói igazolással történik, de lehetnek automatikus módszerek, amelyek minimális emberi beavatkozás nélkül végzik az adathalmazok minőségbecslését. Dinamikus adathalmazok esetén olyan eljárásokra van szükség, amelyek periodikusan újra felmérik a minőséget, és mindezt kellő hatékonysággal teszik.

– Részhalmazok minőségbecslését végző algoritmusok fejlesztése. Adott a teljes adathalmaz, amelynek minőségbecslését már elvégeztük; mit mondhatunk akkor egy részhalmaz (például speciális lekérdezésekre adott válaszok) minőségéről. Ebben az esetben a minőség inhomogenitásából az következik, hogy a válaszok minősége lekérdezésenként változik.

– Adatminőségre vonatkozó információk használata az adattisztítás folyamatában. Ilyen információkkal tovább növelhetők a statisztika eljárásai a hiányzó, hibás vagy zajos adatok kizárására/pótlására. Különösen ott használhatók jól a minőségi információk, ahol az adatok egymást átfedő adatforrásokból származnak.

AZ ADATBÁNYÁSZAT IGÉNYEI

Az adatbányászat műveléséhez először is arra van szükség, hogy összegyűjtsünk annyi adatot, amennyit csak lehetséges. Ha az elérhető adatok papíralapú archívumokban találhatók, akkor ezeket számítógépes adatbázis formájában rögzíteni kell. Ha az adatok már eleve elektronikus formában rögzítettek, akkor a létező adatbázisok újraszervezésére, transzformációjára lehet szükség. Ehhez legtöbbször informatikai szakemberre van szükség. Az így kialakított adatbázis-gyűjteményt adattárháznak nevezzük.

Ezután szükség van egy olyan eszközre, amely összeköti az adattárházat a statisztikai elemzéseket végző szoftverrel. Az összekapcsolás során szükség lehet arra, hogy a kritikus megbízhatóságú vagy hiányos adatokat kivonjuk a vizsgálatból. Néhány adatbányász-szoftver képes az adatok közvetlen kivonására az adattárházból. A memóriaproblémák elkerülése érdekében néhány eszköz nagyon takarékos az adatbehozatal tekintetében és nagyon gyors a párhuzamos feldolgozásnak köszönhetően. (Például az IBM adatbányász-szoftvere, az Intelligent Miner, több száz gigabájtnyi adat elemzésére képes.)

Az adatbányász-szoftver hatékony használatához adatelemzésben jártas szakemberre van szükség, mivel az elemzések értékelése gyakran igen bonyolult. Lényeges továbbá, hogy az alkalmazó tisztában legyen az eszköz működésével, hogy elkerülhetők legyenek a félreérthető következtetések. Ezért a legtöbb forgalmazó konzulenseket biztosít adatbányász-szoftveréhez.

Az adatbányász-szoftvereknek három generációját különböztethetjük meg. Az első generációs szoftverek egyetlen algoritmust valósítottak meg, vagy algoritmusok gyűjteményét tartalmazták vektor értékű adatokban való bányászásra. Ezek a rendszerek már kereskedelmi forgalomban is kaphatók, és gyártóik folyamatosan foglalkoznak továbbfejlesztésükkel.

A második generációs szoftverek már jelentős mértékben túllépték az első generáció által nyújtott szolgáltatásokat, és új funkciókat támogatnak:

– összetettebb adatok kezelése,
– az adatbányászat és az adatkezelés integrációja,
– az adatbányászat integrálása előrejelző modellekkel,
– nagyobb méretű és többdimenziós adathalmazok kezelése,
– adatbányász sémák és parancsnyelv alkalmazása.

A második generáció szoftverei inkább csak kutatóműhelyekben léteznek, kevés kereskedelmi forgalomban kapható szoftver sorolható ebbe a kategóriába. Az új szolgáltatások támogatása további kutatásokat igényel, elterjedésük öt éven belül várható [4].

A harmadik generáció szoftvereire leginkább az jellemző, hogy elosztott adatbázisokban található nagyon heterogén adatokban képesek bányászni. Ezek a szoftverek legtöbbször beágyazott rendszerek, azaz más rendszerek számára közvetlenül szolgáltatják az adatbányászat eredményeit, így magas szintű kommunikációs képességgel kell rendelkezniük. Szoftvertechnikai megoldásként az ún. „intelligens ügynököt" szokták alkalmazni. Kereskedelmi forgalomban való elterjedésükhöz még költséges kutatásokra van szükség.

A következő táblában felsoroljuk a legismertebb adatbányász-szoftvereket annak jelölésével, hogy az ismertetett modellek közül melyeket támogatják. A megadott Internetcímeket felkeresve a szoftverekről részletes információ is nyerhető. Külön megemlítendő a Two Crows Corporation (*http://www.twocrows.com*), amely adatbányász-szoftverek értékelésével foglalkozik.

Adatbányász-szoftverek

Szervezet	Termék	OSZ	REG	IDŐ	KLA	KAP	SOR
ANGOSS Int. Ltd.	KnowledgeSEEKER					X	
(*http://www.angoss.com*)	KnowledgeSTUDIO				X	X	
IBM							
(*http://www.ibm.com*)	The Intelligent Miner	X	X	X	X	X	X
Integral Solutions Ltd.							
(*http://www.isl.co.uk*)	Clementine	X	X	X	X	X	X
Megaputer Intelligence Ltd.							
(*http://www.megaputer.ru*)	Polyanalyst	X	X	X	X	X	
SAS Institute Inc.							
(*http://www.sas.com*)	SAS Enterprise Miner	X	X	X	X	X	
Silicon Graphics Inc.							
(*http://www.sgi.com*)	Mineset	X			X	X	
SPSS Inc.							
(*http://www.spss.com*)	SPSS Products	X	X	X	X	X	
Statsoft							
(*http://statsoft.com*)	STATISTICA	X	X	X	X		
Thinking Machines							
(*http://www.think.com*)	Darwin	X	X				

Megjegyzés. A táblában használt rövidítések a következők: OSZ – osztályozás; REG – regresszió-számítás; IDŐ – idősorelemzés; KLA – klaszterezés; KAP – kapcsolatelemzés; SOR – sorozatkeresés.

*

Az adatbányászat tulajdonképpen a lekérdező és jelentéskészítő rendszerek természetes fejlődéséből jött létre. Aki ma lekérdezéseket hajt végre és jelentéseket állít elő, élvezheti az adatbányászat nyújtotta előnyöket. Mivel az elérhető adathalmazok száma és mérete egyre növekszik, az adatok közvetlen elérése mind lehetetlenebbé válik. Emiatt gyakrabban van szükség elemző eszközök használatára. A számítógépek ma már képesek biztosítani, hogy hetek és hónapok helyett percek és órák alatt előállíthatók az információk nagyméretű adatbázisokból. Mivel az adatbányászat folyamata szisztematikus eljárásokból áll, az adatbányász-eszközök olyan információkat is képesek felfedezni, amelyek egyébként rejtve maradnak. Ezek az információk a piac, illetve a szervezet működésének jobb megismerését eredményezhetik, ezért az adatbányászat alkalmazása versenyelőnyt jelenthet.

Az üzleti élet és a társadalom fontos döntései legtöbbször nagyméretű és összetett adatbázisok elemzésén alapulnak. Ebben a döntéshozatalban az adatbányászat sokat segíthet, amint azt a következő reprezentatív példák is mutatják [10].

– *Marketing*: annak előrejelzése, hogy mely ügyfelek válaszolnak egy postázott reklámanyagra, vagy vásárolnak meg egy terméket; ügyfelek osztályozása demográfiai adatok alapján.

– *Bankügy*: kockázatos kölcsönök és bankkártyával elkövetett visszaélések előrejelzése, új ügyfelek minősítése, új banki szolgáltatások potenciális ügyfeleinek meghatározása.

– *Kereskedelem*: forgalom előrejelzése, helyes raktárkészletek meghatározása, szállítások ütemezése.

– *Termelés*: géphibák előrejelzése, a gyártókapacitás optimális irányítása szempontjából kulcsfontosságú tényezők meghatározása.

– *Részvénykereskedelem*: részvényárfolyam változásainak előrejelzése, az eladás és vásárlás megfelelő időpontjának előrejelzése.

– *Biztosítás*: rizikótényezők meghatározása, biztosítási költségek előrejelzése, új biztosítási szolgáltatások potenciális ügyfeleinek meghatározása.

– *Számítógép-hardver és -szoftver*: lemezmeghajtó hibáinak előrejelzése, potenciális rendszerfeltörések előrejelzése.

– *Kormányzat*: költségvetési források alakulásának előrejelzése.

– *Egészségügy*: kritikus betegségek demográfiai összefüggéseinek meghatározása, betegségek szimptómáinak jobb meghatározása, kezelések pontosabb kiválasztása.

– *Rendőrség*: bűnözési módok, helyek, viselkedések és jellemzők nyomon követése a jobb bűnmegelőzés érdekében.

Az adatbányászatnak ugyanakkor számos problémával kell megküzdenie, amelyek állandó kutatási feladatokat adnak a szakembereknek. Néhány ilyen probléma:

– az adatokat a legritkább esetben hozzák létre, rögzítik és tárolják adatbányászat céljaira;

– a létrejött adathalmaz többdimenziós, a dimenziók száma igen magas lehet, ami mind az adathalmaz megjelenítésében, mind bemutatásában gondot jelent;

– oksági összefüggések megjelenítése: ha kapcsolatot találunk egyes változók vagy minták között, hogyan jellemezhető ez oksági kapcsolatként;

– az adatbányászat zárt világot feltételez: adott adathalmazon belül keresünk összefüggéseket, miközben a valódi befolyásoló tényezőket nem feltétlenül sikerül megragadni;

– az adatbányászat eredményei a felhasznált adathalmazra vonatkoznak, nekünk pedig a valós világgal kapcsolatos megállapításokra van szükségünk: hogyan igazolhatók az eredmények;

– bár a hiányzó adatok/változók figyelembevételére számos eljárást dolgoztak ki, nem állíthatjuk, hogy ezt a problémát maradéktalanul megoldották volna;

– a talált minták hogyan értelmezhetők és egy adatbányász-szoftver hogyan adhat hasznos segítséget az értelmezéshez;

– az automatikus mintakeresés során az adatbányász-eljárások először a nyilvánvaló összefüggéseket találják meg: kérdés, hogy a hatékonyság növelése érdekében miként lehet ezeket kizárni, további kérdés, hogy mikor kell abbahagyni az ismeretkeresést.

A mai adatbányász-eszközök törekednek az automatizálásra, de az eredmények nem automatizálhatók. Az alkalmazók nem mentesülhetnek attól, hogy értsék szervezetük működését, üzletmenetük alakulását. Nem elegendő egy a legújabb statisztikai elemző eljárásokat alkalmazó adatbányász-szoftver beszerzése, ezeket a módszereket érteni és értelmezni is kell. Az adatbányász-szoftvereket kifejlesztő cégek mindent megtesznek az eredmények minél jobb vizuális megjelenítéséért, ma már az animáció alkalmazása sem ritka. Az alkalmazás további problémája, hogy nem minden feltárt összefüggés igényel beavatkozást az üzletmenetbe.

A felsorolt kérdések vizsgálatára ösztönzően hat, hogy az előrejelzések szerint az adattárházak piaca – és ide kell sorolnunk az adatbányász-szoftvereket is – évente 40 százalékkal növekszik, 1998-ra az előrejelzés 8 billió dollárról szólt.

IRODALOM

[1] *Brand, E. – Gerritsen, R.*: The DBMS guide to data mining solutions. DBMS–ONLINE. 1998. évi 7. sz.

[2] *Decker, K. M. – Focardi, S.*: Technology overview: A report on data mining. Technical report CSCS TR-95-02, CSCS-ETH. Swiss Scientific Computing Center. 1995.

[3] *Frawley, W. – Piatetsky-Shapiro, G. – Matheus, C.*: Knowledge discovery in databases: An overview. *AI magazine*. 1992. Ősz. (Fall.) 213–228. old.

[4] *Grossman, R. L.*: Data mining: Challenges and opportunities for data mining during the next decade. Mining and Managing Massive Data Sets ’98 konferencia, 1998. Lajolla. (Kanada) 1998. február 5–6.

[5] *Han, J.*: From database systems to knowledge-base systems: An evolutionary approach. Conference tutorial at the Eleventh International Conference on Data Engineering. Taipei. 1995.

[6] *Han, J.*: Data mining techniques. ACM-SIGMOD ’96 Conference tutorial.1996.

[7] *Holsheimer, M. – Siebes, A.*: Data mining, The search for knowledge in databases. CS-R9406 Report, CWI. Amsterdam. 1994.

[8] *Rocke, D. M.*: A perspective on statistical tools for data mining applications. Mining and Managing Massive Data Sets ’98 konferencia, 1998. Lajolla. (Kanada) 1998. február 5–6.

[9] *Thearling, K.*: From data mining to database marketing. Data Intelligence Group (DIG). White paper. 1995/02. Pilot Software.

[10] *Turban, E. – Aronson, J. E.*: Decision support systems and intelligent systems. 5. ed. Prentice-Hall International. New Jersey. 1998. 890 old.

[11] Two crows corporation: Introduction to data mining and knowledge discovery. 2. ed. Two Crows Corporation. Potomac. 1998. 31 old.

TÁRGYSZÓ: Adatbázis-kezelés. Adatbányászat. Statisztikai információ.

SUMMARY

„Data Mining” has become a buzz-word within the computer industry for extraction of knowledge or information from large databases. This new technology has ideas which have existed in the statistics community for about 15 years under the name of exploratory data analysis. The convergence of these ideas coupled with recent advances in storage technology and database structures offer an interesting, exciting new technology for producing some strategical information. The most advanced products enable several users to work at the same time on the same data from several terminals.