

Adatbányászat a gyakorlatban

A számítógép adatok tömkelegét képes tárolni, de a megszerzett információk hasznosítása már intelligenciát igényel. Az adatok rendezése, rendszerezése, csoportosítása nélkül használhatatlan a nagy mennyiségű tárolt adat. Megfelelő szoftverekre van szükség, hogy el ne merüljünk az információk tengerében. De a legfontosabb az adatok kezelési irányelveinek kidolgozása.

Bevezetés

A vállalatoknál tárolt adatok mennyisége folyamatosan nő. A hardver ára csökken, a berendezések gyorsan elavulnak, a számítógépeket újakra cserélik. Egyre több információ halmozódik fel, ugyanakkor egyre nehezebbé válik azok áttekintése.

Ha egy vevő számítógépen megrendel egy árucikket – pl. egy ruházati cikket egy áruházból –, önmagáról számos értékes információt szolgáltat. Az internetszolgáltató tárolja az időpontot, az online kapcsolat időtartamát, a helyre vonatkozó adatokat, a kapcsolat módját (ISDN, ADSL, analóg, mobil) és 100 000 további ügyfél adataival együtt elemzi. A csomagküldő szolgáltatások adatait is gondosan kezelik, hiszen sok mindent meg lehet tudni a divat iránti érdeklődésről, az árucikkek kombinációiról, a megrendelők nem és életkor szerinti eloszlásáról, a fizetés módjáról, a vásárlás gyakoriságáról, az ügyfél elégedettségéről stb. A rendelés közvetlen hatással van az áruház készletezési rendszerére, ahol pontosan rögzítik mikor, melyik árucikk, milyen áron hagyta el a raktárt.

Szinte elképzelhetetlen mennyiségű adatot tárolnak a különböző rendszerek, különböző formátumokban. Az áruház az adatokat egységes, konszolidált formára kívánja hozni. Rendkívül időigényes a rendelkezésre álló adatok „tisztítása”.

Megfelelő eszközök nélkül nehéz a felhalmozott adatokból információkat, majd a szükséges ismereteket levezetni. Megszületik az „adatbányászat” iránti igény. A bányászat példájával élve: a hatalmas adathegyben kis „információ-göröngyöket” keresnek. A kapott információk, megfelelő módon felhasználva, az aranynál sokszor értékesebbek lehetnek! Az adatbányászatban a legfontosabb, hogy eddig ismeretlen információ kerül a felszínre. Maga az adatbányá

szat azonban még nem biztosít végleges megoldást. Az adatbányászati folyamat módszerének helyes megválasztásán van a hangsúly. Fontos továbbá az adatbányászat eredményeinek integrálása az üzleti folyamatba.

Hol van értelme az adatbányászatnak?

Általánosságban az adatbányászat mindenütt értékes szolgáltatokat tesz, ahol sok az információ és ahol a folyamatokat tökéletesíteni lehet: vagyis a gazdaság csaknem valamennyi területén. Az adatbányászatot a gyakorlatban elsősorban ügyfélelemzésre vagy műszaki folyamatok elemzésére használják fel.

Bankok, biztosítótársaságok, híradástechnika

A hitelintézetek a személyi adatok (pl. kor, nem, foglalkozás, bankszámlahelyzet, további tranzakciók) alapján tudják előre jelezni ügyfeleik hitelképességét. Az elmúlt hetek, hónapok, évek részvényárfolyamaiból levezetett deviza- és kamatindikátorai, a jelenlegi részvényárfolyam, valamint egyéb indikátorok (DAX, dollárárfolyam, kamatok stb.) alapján próbálják előre jelezni a várható részvényárfolyamot.

A hitelkártyával kapcsolatos visszaélések felderítése érdekében az ügyfél régebbi tranzakcióit elemzik, hogy „megismerjék viselkedését” és felismerjék a félrevezető, hamis igényeket.

A pénzügyintézetek és a biztosítótársaságok mindinkább felismerik a „gépi” ügyfélkiválasztás előnyeit. Döntési fák (decision trees) segítségével kiemelik az értékes ügyfeleket és elemzik a viselkedésüket a biztosítási termékek fogyasztása, károkozás és prémium bónusz szempontjából. Így például az egyik vezető svájci járműbiztosító társaság intenzív adatbányászati-elemzés segítségével az alábbi következtetésre jutott: a Saab gépkocsik tulajdonosai a biztosító számára rendkívül értékes ügyfelek, mert nagyon ritkán keverednek közlekedési balesetbe, tehát kevés költséget okoznak. Ezenkívül erősen érvényesül a biztonságtudatuk, általában sok biztosítási terméket, például teljes körű cascót és felelősségbiztosítást vesznek igénybe, a leglojálisabb, leghívebb ügyfeleknek számítanak.

Ipar

Az iparban céltudatos adatbányászattal optimalizáltak termelési folyamatokat. Így például nagy épületek esetében adatbányászattal figyelték az épületben levő felvonók használatát. Az adatelemzés számára az állási helyzetre, a felvonómozgásra, a napszakra és a hét napjára vonatkozó adatokat használták fel. Az eredmények alapján egyedi felvonószabályozást alakítottak ki,

amelyik reggel, este és munkanapokon, a hétvégeken eltérő módon, azonban maximális mértékben személyre optimalva működik.

A rendelések átfutási idejét azzal lehet csökkenteni, hogy a rendelésállomány és a raktárkészletek, valamint a gépkapacitások alapján előrejelzéseket készítenek. Az elemzés a múltira vonatkozó adatok alapján lehetővé teszi a vállalatok számára termékeik várható forgalmának meghatározását. A BioComp System Inc. a forgalmazási előrejelzés figyelembevételével, az adatbányászat alapján fejlesztett ki egy terméket. További ipari alkalmazási lehetőségek: minőségellenőrzés, hibák előrejelzése, műszaki értékek prognózisa. Autóbuszok és repülőgépek szállítási menetrendjét lehet optimalni, amennyiben előrejelzik az utaslétszámot.

Bevásárlóközpontok és csomagküldő szolgálatok

Nagykereskedelmi célokra való felhasználásra példa az ügyfelek összehasonlítása földrajzi és szociológiai kritériumok alapján, a válaszolók arányának előrejelzése, valamint a szállítmányok optimális eljuttatása. Melyek azok a termékek, amelyeket az ügyfelek közösen fogyasztanak el? Milyen esetekben kombinálhatók a „csalogató kínálatok” az új termékekkel? Az egyik élelmiszer-áruház ezzel a módszerrel kísérte meg a heti élelmiszer-vásárlások előrejelzését.

E-kereskedelem

A hálózatok területén is vannak lehetőségek: melyik honlapról térnek át az érdeklődők a cég saját honlapjára? Milyen kereszthivatkozások vezetnek a leghatékonyabban a saját honlapra? Adatbányászattal az ügyfelek viselkedését is jobban lehet felismerni és irányítani. Ha például valaki az Amazon honlapján egy könyvet rendel, a következő honlapon információt kaphat arról, hogy más, hasonló érdeklődési körű ügyfelek milyen könyveket vásároltak korábban.

Egészségügyi ellátás

Adatbányászat segít a különböző betegségek legkedvezőbb kezelési módszereinek meghatározásában. Orvosi diagnózis is kidolgozható neuronhálózatok felhasználásával, a beteg jellemző adatainak ismeretében.

Sportfogadások

Lóverseny- és agárverseny-fogadások. Neuronhálózatok segítségével próbálták előre jelezni a sportesemények eredményeit. Ilyen például a

Greyhound Racing Prediction, amelyik az ID₃ szoftver változatait használja fel.

Adatbányászati módszerek

Elvileg két alapvető adatbányászati módszert különböztetnek meg: feltételezések (hipotézisek) igazolása adatok alapján és feltételezések automatikus kidolgozása az adatok felhasználásával. A hipotézisek igazolásakor analitikai módszerrel egy hipotézist kell felállítani. Ezután kerülhet sor ennek ellenőrzésére a rendelkezésre álló adatok alapján. Ha pl. egy meghatározott folyamat szűk keresztmetszetet okoz, céltudatos módszerrel lehet az ilyen folyamatokat elemezni és optimalni. Jelentős szakértői ismeretekre van szükség az üzleti folyamatokról, ha valaki ezt a módszert akarja igénybe venni. Az adatok közötti összefüggések feltételezéséből kell kiindulni. Sokoldalú módszer a hipotézisek automatikus keresése. Ez az eljárás két további változatra osztható. A közvetlen eljárás esetében megadják az elérendő változót és keresik a függő változókat. A közvetett eljárás estében nem definiálnak előre változót, csupán megkísérik, hogy a tárolt adatok alapján megtalálják az összefüggéseket, illetőleg korrelációkat. Ezt a módszert lehet többek között az ügyfélállomány csoportosítására felhasználni.

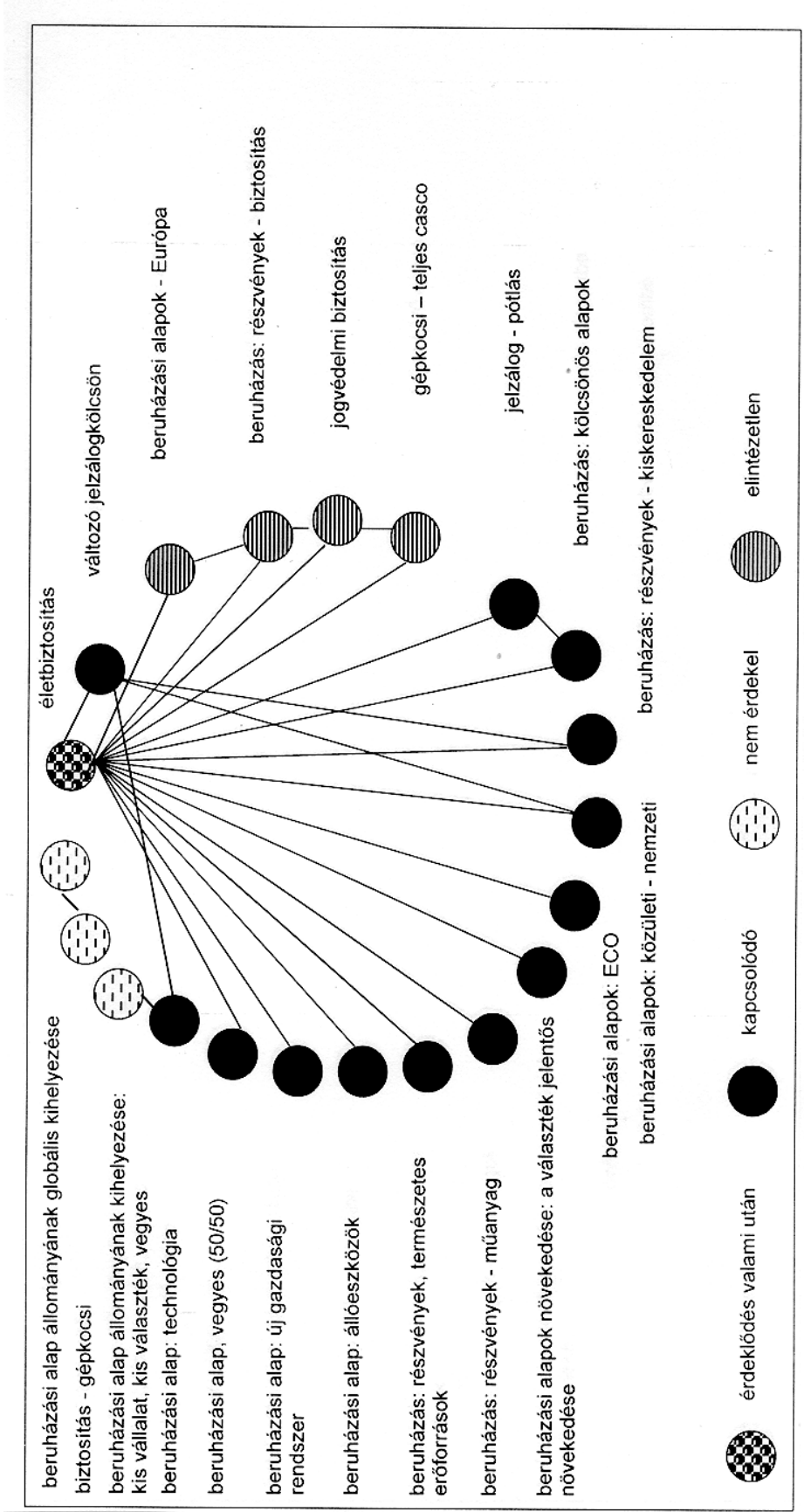
Adatbányászati eljárások

Mindegyik eljárás egy meghatározott problématerület estében tesz hasznos szolgálatot. A felsorolt eljárásokat elsősorban kereskedelmi célokra alkalmazzák.

Árukosárelemzés

Az árukosárelemzés (1. ábra) a cluster-elemzés csoportjába tartozik.

A közösen eladott termékek csoportjait keresik. Kizárólag a kiskereskedelembe kerülnek felhasználásra és a lehető legkedvezőbb lehetőséget biztosítják a vásárlási viselkedés elemzéséhez. A kapott eredmények segítségével lehet az egyes termékeket csoport-hozzá tartozásuk alapján a polcokon elhelyezni. Van-e lehetőség az árukosárelemzést az ügyfelekre vonatkozó adatokkal kapcsolatba hozni? Ki lehet-e számítani a jövőben bekövetkező vásárlások esetére a bevásárlási valószínűséget?

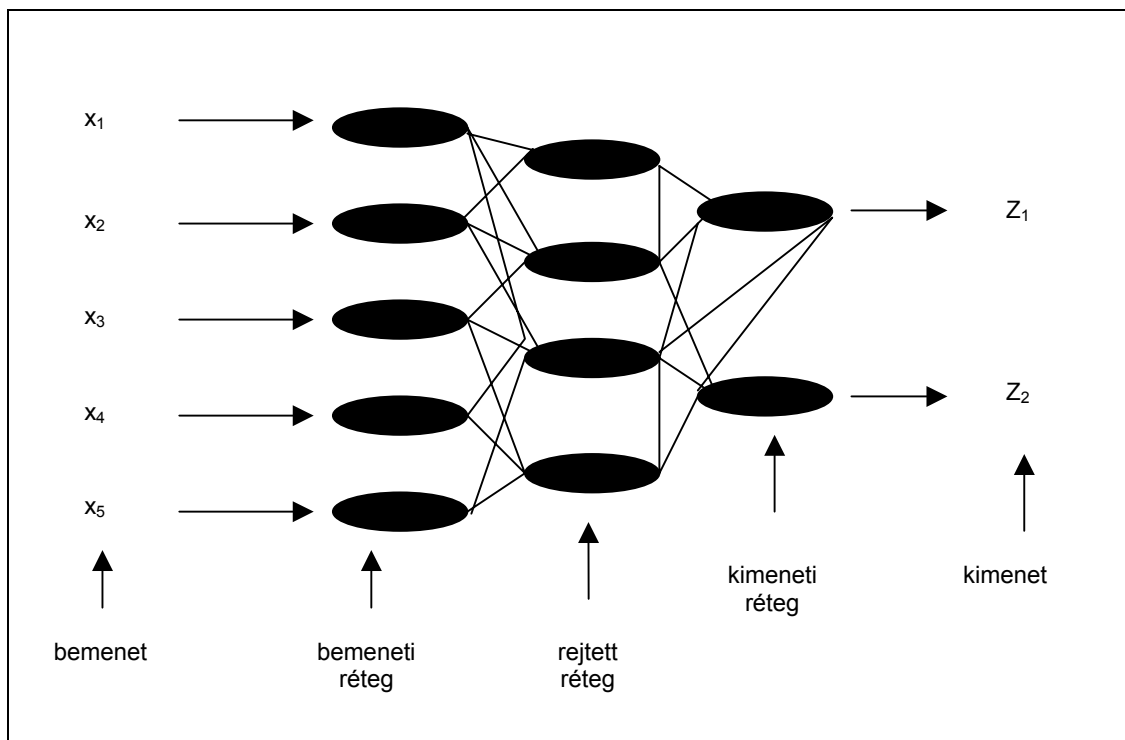


1. ábra Árukösárelemzés

Az egyik nagy áruházlánc az adatbányászat alapján megállapította, hogy az üdvözlőlapokat és az illatszereket a vevők gyakran együtt vásárolták. Ezt a két terméket egymás mellett helyezték el. Ennek felismerése révén, az árukosárelemzés eredményeit kihasználva, a két cikk kereskedelmi forgalmát egyharmaddal növelték. További példa a pénzügyekben a biztosítások összekapcsolása további pénzbefektetésekkel, pl. jelzálogkölcsönökkel.

Következtetés esettanulmányok alapján

Ezzel az eljárással lehet a jövő döntéseit a múlt tapasztalatai alapján levezetni. Ennek érdekében az egyes esetek jellegzetes tulajdonságait, a döntés szempontjából lényeges paramétereket és a megkötött üzlet végeredményét adatbankban tárolják. Új döntés előtt a paramétereket értékelik és összehasonlítják az adatbankban szereplő adatokkal. Az adategyezés mértékével arányos az előrejelzés pontossága.



2. ábra Neuronhálózatok

Neuronhálózatok

A neuronhálózatok (2. ábra) olyan információfeldolgozó rendszerek, amelyek sok egyszerű egységből, neuronból állnak. Az információkat szimulá

cióval, irányított kapcsolaton keresztül cserélik. Ahhoz, hogy egy neuronhálózat ésszerűen legyen használható, a feladatra be kell gyakoroltatni. Az alapismeretek elsajátítása a bemeneti és a meghatározandó értékek megadásával történik. Értelemszerűen kétféle tanulástípust lehet megkülönböztetni.

Az ellenőrzött tanulást gyakran arra használják fel, hogy az adatokat osztályozzák és előre jelezzék. Így például lehetőség van arra, hogy tetszőleges időpontra meghatározzák a forgalom alakulását. Az ilyen neuronhálózatot a múltból származó, már ellenőrzött példák alapján lehet betanítani. Vagyis, régebbi üzleti helyzeteket és ezeknek a forgalomra gyakorolt hatását használják fel kiindulópontként.

Az ellenőrizetlen tanulást adatok csoportosításához, célcsoportok elemzéséhez használják fel. A célcsoport elemzésekor a neuronhálózat az ügyfelek adatai között minta után kutat és a talált minta alapján az ügyfeleket célcsoportokra osztja.

A neuronhálózatok esetében az jelenti a legnagyobb problémát, hogy a megoldáshoz vezető út nem áttekinthető. Az eredmények azonban kitűnőek. A NASA a neuronhálózatokat az űrrepülőgépek karbantartásához használja fel. A módszer felhasználási területe csaknem határtalan.

Halmazok automatikus elemzése

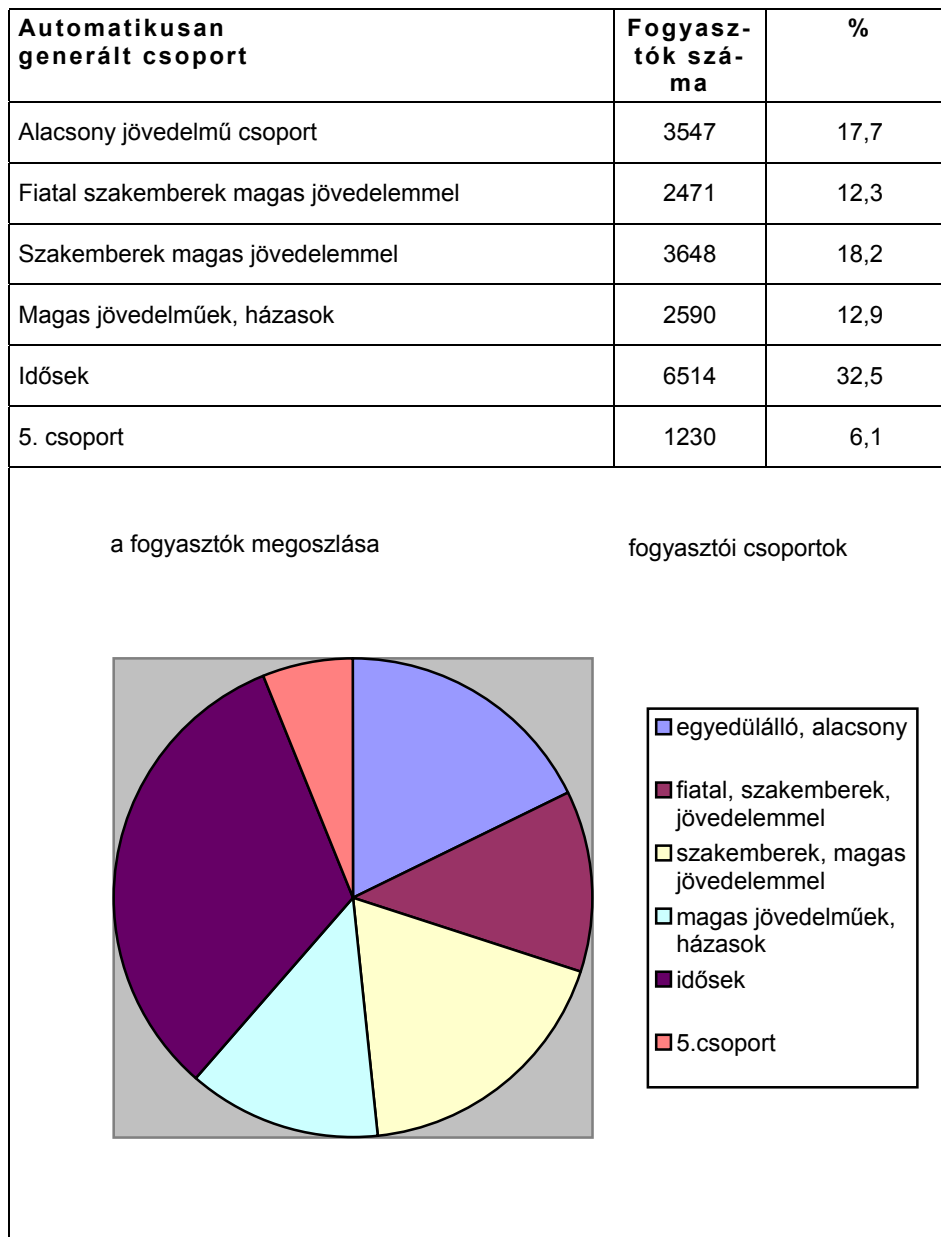
A halmazok automatikus elemzésének módszerét adatkészlet-csoportokat kiválasztva használják. Az elemzés célja, hogy a nyers adathalmazból még ismeretlen összefüggéseket találjanak. Legtöbbször igen nagy adatállományon belül, első lépésként, a hasonló tulajdonságú adatokat kiemelik, majd más eljárásokkal folytatják azok vizsgálatát.

Adatkészletek közötti kapcsolatok elemzése

Ezzel a módszerrel kísérelik meg az egyes adatkészletek közötti kapcsolatokat feltárni. A cél általában, hogy a marketingtevékenységet jobban az egyes ügyfelekre koncentrálják. Így például lehetőség van arra, hogy a háztartásokat jövedelmi szintjük alapján szőlítsák meg (3. ábra).

Valamilyen új termék bevezetésének kezdetén, igen gyakran, az ügyfelek csoportosítása a halmazelemzés módszerével történik. Minél pontosabban jelölhető ki valamelyik ügyfélcsoport, annál olcsóbban vezethető végig a marketingkampány. Gyakran előfordul, hogy a meglévő ügyfél-adatállományt három, A, B és C ügyfélcsoportra osztják. az A jelenti az értékes, a B a potenciális, a C a gyengébb ügyfeleket. Az ilyen csoportosítás folyamatosan egy dinamikus tényezőt eredményez. Ha a csoportosítást időszakosan végzik, akkor dinamikus csoportosításról lehet beszélni. A cél az, hogy a B ügyfeleket át le

hessen vezetni az A ügyfelek közé, azaz a potenciális ügyfelekből értékes ügyfelek váljanak.



3. ábra Rendezetlen halmazok elemzése

Adatbányászati szoftverek

Az adatbányászat szempontjából három kritérium szerint lehet a szoftvermegoldásokat megkülönböztetni:

- Egy számítógépen futó (desktop) megoldások, amelyek statisztikai és adatbányászati értékelések céljait szolgálják.

- Ügyfél/szolgáltató (client/server) alapú rendszerek, amelyek elosztott számítástechnikai kapacitás-igénybevételt tesznek lehetővé és statisztikára, valamint adatbányászatra vannak szakosítva. Ezek a rendszerek a múltban gyakran egyedi megoldásokat képviseltek, azonban egyre inkább valamennyi közismert platform esetében felhasználhatók.
- Java nyelven programozott, ügyfél/szolgáltató rendszerek, amelyek messzemenően platformfüggetlenek. Ezek teljesítőképessége az elosztott számítástechnikai teljesítmény következtében igen nagy. Statisztikára és adatbányászatra vannak szakosítva. Ezenkívül ETL folyamatok (kiemelés, átalakítás, betöltés) számára is lehetőséget biztosítanak. Ezen túlmenően, zökkenőmentesen integrálhatók a meglévő számítástechnikai architektúrákba.

A vállalati adatok dinamikus elemzésére a jövőben csak az utóbbi kategóriába sorolható szoftvermegoldások kerülhetnek szóba. A gyors döntés szempontjából mérvado információkat igen gyakran dinamikus elemzik. A heti, a napi vagy az online adatbányászati elemzés szempontjából (hálózati adatbányászat) fontos, hogy a teljes ismeretkimutatási folyamat egyetlen szoftvermegoldáson belül legyen megvalósítható.

Az ismeretkimutatási adatbányászás (Knowledge-Discovery-Data-Mining-Prozess, KDDM) az alábbi lépésekből tevődik össze: az adatok kiválasztása az operatív rendszerből, majd ezek előzetes feldolgozása, ami az adatok konszolidálását valósítja meg. Ezután az adatok átalakítása a kívánt adatbányászati formátumra. Az algoritmusok segítségével lehetővé válik a tulajdonképeni adatbányászati elemzés. Ezt követi az adatok értelmezése, valamint a kapott adatok kezelése. Az utolsó jelentős lépés az „adatdúsítás”, ami az eredményeket visszajuttatja az operatív rendszerbe.

(Dr. Barna Györgyné)

Grubenmann, U.: Mehr Licht im Datenschungel. = New Management, 70. k. 11. sz. 2001. p. 62–68.

Jakob, R.: Datawarehousing: Es kommt drauf an, was man drauf macht. = New Management, 71. k. 6. sz. 2002. p. 68–74.