Adatbányászat és adatvédelem

Stifán Orsolya*

Budapesti Műszaki és Gazdaságtudományi Egyetem
Gazdaság- és Társadalomtudományi Kar
Információ- és Tudásmenedzsment Tanszék
1111 Budapest, Sztoczek u. 2. St. ép. I. em. 117.
Telefon: (36 1) 463-1832, Fax: (36 1) 463-4035
e-mail: stifano@t-mobile.hu

Absztrakt

Bár a szakirodalomban egyre több szó esik az adatbányászat területéről, gyakorlati alkalmazásával Magyarországon csak a legnagyobb vállalatok, a bankok, biztosítók és telekommunikációs cégek esetében találkozhatunk. Az 1992. évi LXIII. törvény foglalkozik a személyes adatok védelmével, melynek az EU-csatlakozás utáni hatályos szövegében már megjelenik az automatizált egyedi döntés fogalma, még eléggé puhán megragadva az adatbányászat egy alkalmazását, de a törvényben kétségkívül felfedezhetjük a szabályozás első jeleit. A tanulmány az adatbányászat és az adatvédelem viszonyát mutatja be, kiemelve a kritikus pontokat.

Kulcsszavak: *adatvédelem, adatbányászat, személyes adat, PPDM (Privacy preserving data mining)*

1. Az adatbányászat alapjairól

1.1. Az adatbányászat mibenléte

Az adatok gyűjtésének célja, hogy azokból információt nyerjünk ki, mely felhasználható üzleti döntések eredményes meghozatalához, azaz „rejtett, ismeretlen, potenciálisan hasznos tudás kinyerése az adatokból, nem triviális módon”. [1] Az adatbányászat ennek a folyamatnak egy lépése, melyet megelőz az adatkiválasztás, tisztítás, bővítés és kódolás. A szó maga először negatív felhanggal jelent meg az 1960-as években, lenézett tevékenységhez hasonlítva azt, hiszen a korábbi alapos statisztikai módszerekkel szemben az adatbányászat valóban kevésbé „tűnt” tudományos tevékenységnek. A bányászat arra utalt, hogy ha valaki kellő időt tölt az adatok vizsgálatával, akkor valószínűleg találni fog

* Stifán Orsolya, III. évfolyamos PhD-hallgató, BME GTK Információ- és Tudásmenedzsment Tanszék.

olyan összefüggést, melynek különös fontosságot tulajdoníthat. Később, a bányászat szó már másra utalt: a valódi bányászattal analóg módon a megmozgatott tömeg és a talált érték viszonyára. Az adatbányászat számtalan definíciójának közös elemei: nagy adatbázisokból, rejtett tudás kinyerése, új, nem várt minták automatikus felfedezése.

Az adatbányászat térnyerését a nagy adattömegek rendelkezésre állása katalizálta, amelyek a vállalatok tranzakciós rendszereiben milliószámra termelődtek napról napra. Később, a hasznosnak vélt adatokat a vállalatok adattárházakba rendezték. Már a 80-as évektől ODBC kapcsolatokon keresztül SQL lekérdező nyelv segítségével sikerrel hozzáfértek az adatok nagy részéhez, de a rejtett információ kinyerésére — melyhez való hozzáférés formalizáltan nem volt megadható — még várni kellett. A multiprocesszoros gépek, a hálózatok elterjedése és a fejlett algoritmusok rendelkezésre állása újabb lendületet adott az adatbányászatnak. A korábbi rendszerekkel (döntéstámogató és felsővezetői információrendszerek) ellentétben — melyek főleg a múltra irányultak — az adatbányász eszközök már a jövőt fürkészték.

1.2. Adatbányászat, OLAP és statisztika

Az adatbányászat mibenlétét sokszor az OLAP-hoz (On-Line Analitical Processing) képest kellett bemutatni, hiszen sok vállalat már alkalmazta ezt a humán intelligencián alapuló riportoló eszközt, amely előre definiált riportokat képes prezentálni. Ezzel szemben az adatbányászat adatvezérelt, emberi előfeltételezések nélkül képes automatizált mintafeltárásra, előrejelzésre. Az OLAP és az adatbányászat egymás segítségére is lehet, erről bővebben lásd [2].

A statisztika az adatbányászathoz hasonlóan szintén modelleket állít fel. Ezek a modellek azonban az ún. top-down módszert alkalmazzák, azaz egy előre megfogalmazott teórián nyugvó hipotézis tesztelése áll a középpontban, adott mintából következtetünk a sokaságra, a teljes alapsokaságra vonatkozó *a priori* ismeretek és egyéb segédinformációk felhasználásával. (A két terület hasonlóságairól és eltéréseiről lásd [3].)

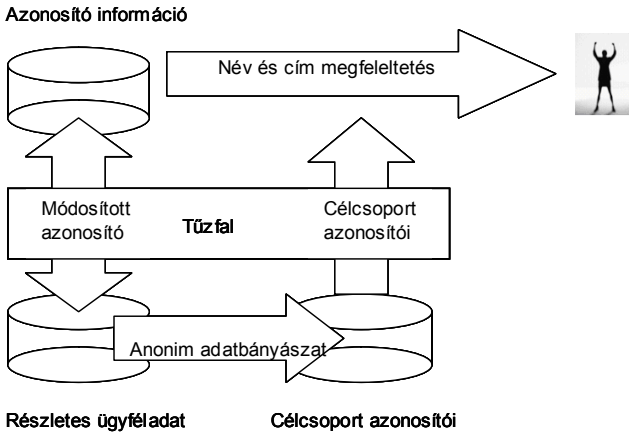
A statisztika és az adatbányászat egymással erőteljesen átfedő területek, nem lehet éles határt húzni a kettő között. Tekintsük például a többváltozós statisztikai modelleket. Ezek először a statisztikai irodalomban jelentek meg, később az adatbányászat is hozzáadott a területhez, és az adatbányászati eszköztárnak is jelentős részét képezik.

2. Adatbányászat és adattárház

2.1. Adattárház és azonosítás

A nagy adatmennyiség rendelkezésre állása szükségessé tette, hogy azokat értelmes struktúrákba rendezzék. Az adattárházak létrejöttének célja, hogy támogassák a vállalatok döntéshozóit az abból nyerhető adatok információvá való transzformálása révén. Jellemzőjük a tematikusság és az integráltság. Előbbi arra utal, hogy témánként rendszerezve találhatók meg benne az adatok; és nem az összes, amely a tranzakciós rendszerekből kinyerhető, csak az, ami fontos lehet. Az integráltság azt jelenti, hogy az adattárház általában több adatbázisból, több forrásrendszerből vagy adatpiacból épül fel. A sikeres adatbányászatnak nem előfeltétele az adattárház megléte, de nagyban segítheti az elemzői munkát.

Az adattárházakban az ügyféladatokhoz való illetéktelen hozzájutást megelőzendő, Berson, Smith és Thearling [6] azt javasolja, hogy az ügyfelekhez rendeljünk egy véletlenszerűen generált számot, mely átveszi az azonosítás funkcióját (általa összekapcsolhatók a különböző táblák), és a tényleges ügyféladatokat az elemzők ne lássák. Fontos, hogy ez az azonosító valóban véletlenszerű legyen, és ne például az irányítószám és a születési dátum felhasználásával generált számot használjunk.



1. ábra: Anonim architektúra [6]

Egy másik egyszerű, ámbár jelentős korlátokat felállító módszer, hogy ne ügyfélszintű adatokat tároljunk az adattárházban, hanem valamilyen hasznos szintre aggregált adatokat. Például ha a marketingtevékenységet az ügyfelek korára és nemére vonatkozó adatok szerint szeretnénk minősíteni, akkor egy egymillió ügyfél adatait tartalmazó adatbázis esetén kb. négyezer aggregátumhoz jutunk (120 korosztály és 2 nem osztály szorzatával

osztva az egymilliót). Ezen a nagyságon még lehet adatbányászati tevékenységet végezni, bár sok egyedi szintű információtól, mintától megfosztottuk magunkat az aggregálás miatt. Szintén nehéz feladat, hogy az összes lehetséges változóosztályra nézve szorzatokat képezzünk. Nem is beszélve a folytonos változókról: hogyan állapítsuk meg az osztályközöket? (Hasonló problémával találkozhatunk OLAP kocka építésekor is.) Olyan eset is előfordulhat, hogy egyetlen ügyfél kerül egy osztályba (például lehet egy 120 éves férfi ügyfelünk), ekkor nem aggregált, hanem egyedi szintű adataink lesznek róla, vagy egyes esetekben akár néhány változó (3-4) segítségével is jól beazonosítható lesz valamely egyén (például irányítószám, gépkocsi típus, életkor).

Az azonosítás kérdésével az Európai Parlament és a Tanács 95/46 EK irányelve is foglalkozik, ám például a nemzeti azonosító kérdésének szabályozását a tagállamokra bízza. [5] Az első magyar adatvédelmi törvényeknek a célja az volt, hogy megelőzze az állam információs hatalmának elterebélyesedését. “A személyi szám elterjedt használata esetén a magánszféra megszűnik, mert a legtávolabb eső különböző célú nyilvántartásokból összehozott adatokból előállítható az ún. személyiségprofil, az érintett tetszőlegesen széles tevékenységi körére kiterjedő és intimszférájába is behatoló művi kép, amely ugyanakkor az adatok kontextusból kiragadott volta miatt nagy valószínűséggel torz is. [...] A nagy mennyiségű összekapcsolt adat, amelyről az érintett legtöbbször nem is tud, kiszolgáltatja az érintettet, egyenlőtlen kommunikációs helyzeteket hoz létre. Megalázó az olyan helyzet, és lehetetlenné teszi a szabad döntést, amelyben az egyik fél nem tudhatja, hogy partnere milyen információkkal rendelkezik róla.” (Az Alkotmánybíróság 15/1991. (IV. 13.) AB határozata; idézi [7])

Az Alkotmánybíróság észlelte, hogy a személyiségprofil felállítása a mai technológiáknak köszönhetően már a személyi azonosító nélkül is lehetséges, de a nagytömegű adatgyűjtést nem tartotta reálisnak az azzal járó kapacitáshiány illetve anyagi áldozat miatt. “A személyi szám a személyes adatok megbízható összekapcsolásának — az adatfeldolgozás mai módjait tekintve — technikai szempontból legelőnyösebb eszköze. Személyes adatokat természetesen össze lehet kapcsolni névvel, és szükség szerint kiegészítő azonosítók segítségével, mint például az anya neve, vagy lakcím. A mai számítógép kapacitások mellett ezek terjedelme sem jelent különösebb problémát. A "természetes" adatok azonban változhatnak (például a név férjhez menéssel vagy névváltoztatással), s előfordulhat, hogy a megkülönböztetéshez további adatok szükségesek; továbbá változó adatok esetén (mint a lakcím) az adatok követése és karbantartása szükséges. Az ezzel járó nehézségek és kiadások jelentős tételként jönnek számba az adatfeldolgozás költség/haszon elemzésénél, s [...] természetes fékjét képezik az indokolatlan adatgyűjtésnek, amire a kéznél lévő személyi szám késztet." (Idézi [7])

Már régen nem ez a helyzet. A kapacitáshiány kérdése megoldható, egyre kisebb ráfordítással. Az algoritmusok pedig valóban lehetővé teszik a személyiségprofil elkészítését, amit ma már nem az állam kíván a legtöbbször alkalmazni, hanem a vállalatok a sikeres marketingtevékenységük végrehajtásához.

A nyilvánosan hozzáférhető adatforrások is elárulhatnak valamit az abban fellelhető adatokról. Például a telefonkönyvben szereplő nevek és számok a vállalati osztályok nagyságáról, és ezek változása esetén akár a stratégiájáról is sok mindent elárulhatnak. Ezt kivédendő, hamis bejegyzéseket tehetünk a telefonkönyvbe, amivel a telefonkönyv célját (a tényleges személyek tényleges telefonszámaihoz való hozzájutást) nem veszélyeztetjük. [8]

2.2. Az adatok elérése

Az adattárház rengeteg adatot tartalmazhat, akár évekre visszamenőleg is, egyéni, azaz ügyfélszinten. Ez elemzési szempontból nagyon hasznos, hiszen az algoritmusok egy döntő hányada a múltbeli események alapján jelez előre, általában minél több a tapasztalati adat, annál pontosabban tudjuk felvázolni a jövőt. Vagy gondoljunk a vállalatok által sűrűn alkalmazott hűségpontokra, hűségszerződésekre. Ha egy ügyfél hosszú időre elkötelezi magát — nagy forgalmat generálva — egy vállalat mellett, annak cserébe valamilyen ajándékot, vagy kedvezményt nyújt a cég. Minden vállalat egészen pontosan kíváncsi a kedvezményezett ügyfelek múltjára, hogy testre szabott ajánlatokkal „kedveskedjen” nekik. De például egy távközlési vállalatnál fontosabb okok miatt is szükség lehet a részletes ügyféladatokra: a múltbeli hívási szokások alapján jó valószínűséggel megbecsülhető a jövőbeli viselkedés, ami a hálózati kapacitás tervezéshez alapvető információ.

A forgalmi adatok tárolásával kapcsolatban az EU 2002/58/EK irányelve még nem foglalkozik részletesen. „Nyolc tagállam fogadott el olyan szabályozást, amely a forgalmi adatok megőrzésének maximumát három hónap és egy év között határozza meg. Ebben a tekintetben azonban uniformizált EU-szabályozás is elképzelhető, amely hosszabb, tizenkét vagy huszonnégy hónapon át tartó megőrzést tenne lehetővé. […] A forgalmi adatok ilyen hosszú megőrzésének lehetőségét számos adatvédő és EU-intézmény kritizálja. Az EU adatvédelmi munkacsoportja 2003 januárjában azt javasolta, hogy az elektronikus kommunikációval kapcsolatban keletkező forgalmi adatokat legfeljebb három vagy hat hónapig lehessen csak tárolni. A 2002-ben megtartott cardiff-i nemzetközi adatvédelmi biztosi konferencián az európai biztosok megállapították, hogy a forgalmi adatok egy évig vagy tovább történő megőrzése ellentétes a célhozkötöttség elvével, ennélfogva elfogadhatatlan. A Privacy International és a Covington & Burling nevű nemzetközi ügyvédi iroda nemrégen közzétett állásfoglalása szerint az adatok ilyen hosszú megőrzése

ellentétes az Európai Emberi Jogi Egyezmény 8. cikkével és a strasbourgi bíróság vonatkozó esetjogával.” [9]

3. Az algoritmusokról

3.1. Példa: credit scoring

Az automatizált egyedi döntés megjelenése a törvényekben nagy valószínűség szerint az elterjedt hitelképesség-ellenőrzési gyakorlatra vezethető vissza. A credit scoring rendszerek — mely kifejezést adósminősítésnek, ügyfélminősítésnek vagy hitelminősítésnek fordítják a magyarban — tulajdonképpen egy valószínűséget határoznak meg, annak a valószínűségét, hogy egy hitelkérelmező mekkora valószínűséggel hoz nyereséget vagy veszteséget a vállalat (bank) számára. Amióta lehetőség nyílt a számítógépek kapacitásigényének széles körű felhasználására, azóta a credit scoring is könnyebbé, gyorsabbá és olcsóbbá vált. Az sem elhanyagolandó, hogy a szubjektivitás, amit korábban a hitelügyintéző személye jelentett, teljes mértékben kiküszöbölhető. Nem fordulhat elő, hogy egy személy két különböző bankfiókba betérvén eltérő hitelminősítést kapjon csak azért, mert más személy végezte a hitelbírálatot.

1958-ban jelent meg az első credit scoring rendszer, Bill Fair és Earl Isaac készítette el az American Investment számára. Korábban nem létezett egységes, átlátható és objektív rendszer. Az 1960-as években már olyannyira megnőtt a hitelkérelmezők száma az USA-ban, hogy az egyedi elbírálást nem lehetett fenntartani, egy automatizált rendszerre volt igény. Ez született meg az 1950-es évek végén. Amint észrevették előnyét, gyorsan el is terjedhetett. [11] Credit scoring során többféle statisztikai módszert alkalmazhatunk: logisztikus regressziót, diszkriminancia analízist, neurális hálózatokat, genetikus algoritmusokat, lineáris programozást, stb.

Sok esetben azonban a hitelkérelmezők nem tudják, hogy hogyan használják fel a róluk összegyűjtött információkat, és bizalmatlanok a bankokkal szemben. Néhány hitelintézet, főleg az Egyesült Államokban különösen odafigyel, hogy kellőképpen tájékoztassa az ügyfeleket, hitelkérelmezőket. Az USA-ban széleskörű adatgyűjtésre van lehetőség, a hitelintézetek általában rendelkeznek a következő adatokkal: az igénylő életkora, neme, családi állapota, állampolgársága, végzettsége, gyermekeinek száma, állása, jövedelme, stb. Azonban a kezük meg is van kötve: nem használhatják fel például a fajra, vallásra, nemzetiségre vonatkozó scorecardokat. Az életkor változó pedig mindaddig felhasználható az elemzésekben, amíg az a 62 évesnél idősebb személyeket nem diszkriminálja. [11]

Az Amerikai Fogyasztók Egyesületének (Consumer Federation of America) egyik felmérésében [13] megfogalmazták azt a hat legfontosabb tényezőt, mely a leginkább foglalkoztatja a hiteligénylőket a credit scoring rendszerekkel kapcsolatban. Ezek a következők voltak:

- *Gyorsaság:* a credit scoring rendszerek jóvoltából a hiteligénylés folyamata felgyorsult. Ez a gyors ügyintézés azt az érzetet kelti az ügyfelekben, hogy nem foglalkoznak velük kellőképpen, és kétségbe vonják a rendszer megbízhatóságát.
- *Ügyfélre szabott árazás:* egyre nagyobb az igény arra, hogy személyre szabottan történjen a hitelbírálat, a kisebb kockázatú ügyfelek kisebb kamatot szeretnének fizetni, a kiesett profitot a banknak a nagyobb kockázatú ügyfeleken kell behajtania. Azonban nagyon sokan nincsenek tisztában saját hitelképességükkel, és nem is akarják megismerni a rendszert, így viszont attól tartanak, hogy egy átlagos szintű kamatot fizettetnek velük, vagy feltételezik, hogy a rendszer alapvetően hibás, ezért bizalmatlanok.
- *Diszkrimináció:* pontosan a rendszer automatizáltságából kifolyólag csökken a diszkrimináció veszélye, hiszen a szubjektív elemek minimálisra csökkennek. Azonban adathiány esetén adatpótlásnál figyelni kell arra, hogy az ne okozzon „mesterséges diszkriminációt". Erre hivatalos szervek is odafigyelnek.
- *Statisztikai validitás:* a credit scoring rendszerek nagyrészt statisztikai modellek is egyben, tehát hibával dolgoznak. Átlagban jól teljesítenek, de egyéni szinten előfordulhatnak ezek a hibák. Ekörül mindig is viták folytak, azonban ez a módszer velejárója.
- *Teszteletlen scoring formulák:* a tényleges formulák továbbra is fekete dobozt alkotnak, de nem csak a hiteligénylő számára, de még a hatóságok felé sem publikálják azokat. Néhány amerikai bank ad némi felvilágosítást, hogy az egyes vizsgált ismérvek mekkora súllyal kerülnek beszámításra hiteligényléskor.
- *Nem megfelelő credit riportok:* leginkább az adatok helyességét vonják kétségbe, ami alapján a credit scoring modellek felépülnek. A kétségek nem megalapozatlanok, sok kutatás alátámasztotta ezt.

A rendszerrel szemben az amerikai állampolgárok sokkal inkább ki vannak szolgáltatva, hiszen az Egyesült Államokban az információáramlás szabadsága megelőzi a magánélet védelmét, és a hiányos állami szabályozást kiegészítendő jelennek meg az ipari önszabályozási modellek.

Az adósminősítés nem csak a bankok gyakorlatából ismert, de a távközlési vállalatok is igyekeznek a pénzügyi kockázatot csökkenteni.

„A társaság [Vodafone] a hitelképesség ellenőrzése során az ügyfél által szolgáltatott adatokat egy — Amerikában és Nyugat-Európában alkalmazott, tapasztalati adatokat figyelembe vevő — számítógépes programmal elemzi. Az esetleges elutasítást követően az előfizető tájékoztatást kap arról, hogy lehetősége van „Előre fizető Előfizető"-ként szerződni, illetve a döntés ellen az ügyfélszolgálathoz fordulni annak felülbírálata végett." [10] A Vodafone 1999 novemberében indította el kereskedelmi szolgáltatását Magyarországon, így 2000-ben a vállalat nem sok tapasztalati adattal rendelkezhetett, azaz valószínűleg a nyugat-európai összefüggéseket adaptáltak a magyar piacra, nyilván szakértői korrekcióval módosítva azt. A Pannon GSM általános szerződési feltételeiben is megjelenik az automatizált egyedi döntés, melynek célja „az Előfizető felsorolt adatainak kizárólag számítástechnikai eszközökkel történő értékelése." [12]

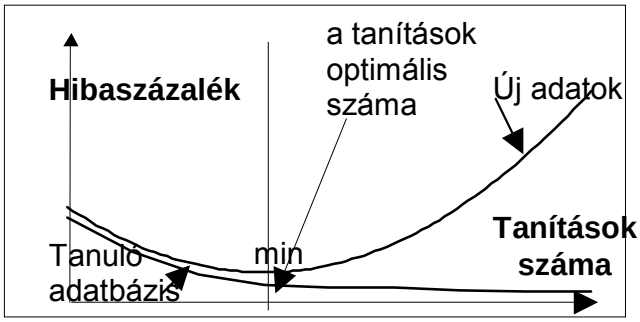
A személyes adatok védelméről és a közérdekű adatok nyilvánosságáról szóló 1992. évi LXIII. törvény röviden érinti az automatizált egyedi döntést is. „9/A. § (1) Kizárólag számítástechnikai eszközzel végrehajtott automatizált adatfeldolgozással az érintett személyes jellemzőinek értékelésére csak akkor kerülhet sor, ha ahhoz kifejezetten hozzájárult, vagy azt törvény lehetővé teszi. Az érintettnek álláspontja kifejtésére lehetőséget kell biztosítani. (2) Az automatizált adatfeldolgozás esetén az érintettet - kérelmére - tájékoztatni kell az alkalmazott matematikai módszerről és annak lényegéről." [14]

3.2. Az algoritmusok működése

Irányított tudásfeltárás során egy előzetesen kiválasztott célváltozót szeretnénk megmagyarázni a többi változó segítségével. (Nem irányított esetben nem határozunk meg célváltozót.) Berry-Linoff [2] a következőképpen ragadja meg az irányított tudásfeltárás lényegét:

1. a még nem osztályozott adatok forrásának beazonosítása;
2. adatelőkészítés;
3. a megfelelő adatfeltáró technikák kiválasztása;
4. az adatok tanuló, teszt és kiértékelő részekre történő felosztása;
5. tanítás, majd tesztelés után módosítás;
6. a modell kiértékelése, ami által vagy megerősítjük, vagy elutasítjuk azt;
7. cselekvés az eredmények hatására;
8. a folyamat újrakezdése a cselekvés hatására előálló adatokon.

A tanítás előtt a rendelkezésre álló adatbázist minimum kettő, de általában három részre osztják fel. Az első szett lesz az, amin az algoritmust tanítjuk, felépítjük vele az alapmodellt, ami feltételezésünk szerint jól leírja a vizsgált sokaságot vagy folyamatot. A teszt adatbázist azért használják, hogy azon kipróbálják a létrehozott modellt, és kalibrálják azt. Ez azért fontos, mert előfordulhat, hogy a tanítás olyan jól „sikerül", hogy bár a tanuló szettre tökéletes modellt hozunk létre, de az annak a szettnek az egyedi sajátosságait tükrözi, ami eltér a sokaságétól. Azaz túl speciális lett a modell, ezt nevezik túlillesztésnek. A kiértékelő szettel a modell hatékonyságát teszteljük, hogy az hogyan teljesít új adatokon. Ezután már csak ki kell választani a felépített modellek közül a legjobbat. A következő ábra mutatja a tanítások optimális számát.



2. ábra: A tanítások optimális száma

Vannak, akik pontosan a tanítási folyamat miatt aggodalmaskodnak. [15] A tanuló adatbázis meghatározása nem triviális kérdés. Kik kerüljenek a tanuló adatbázisba? Hogyan biztosítható, hogy az elemző a tanuló adatbázisból tudomására jutott személyes adatokkal és összefüggésekkel ne éljen vissza? (A legegyszerűbb megoldás a már leírt módosított azonosítók létrehozása, amit viszont az egyik legkevésbé megbízható megoldásnak tartanak, mert az adott vállalatra bízza a módosított azonosító használatának szabályozását, azaz nem igazán ellenőrizhető.) Ha az adatokat módosítjuk, az pedig az algoritmus működését befolyásolná hátrányosan. Egyesek azt is fontosnak tartják, hogy a tanítások optimális számát a fent bemutatott módon határozzák meg a cégek, hogy a modell hasonló hibaszázalékkal dolgozzon új adatok és a tanuló adatbázis esetén, hiszen ha új adatokon nagyobb a hibaszázalék, esetleg megalapozatlanul utasítanak el egy hiteligénylőt.

4. Adatvédelem

4.1. Első megközelítés

A külföldi szakirodalomban gyakran találkozunk a privacy kifejezéssel. Ez a fogalom a személyes adatok védelménél tágabb területet fed le, kiterjed a magánélet egyéb területeire is. A nemzetközi gyakorlatban azonban a privacy-t egyre inkább az „információs privacy" értelmében, a személyes adatok védelmének szinonimájaként használják. [5] Adatvédelem a „személyes adatok gyűjtésének, feldolgozásának és felhasználásának korlátozását, az érintett személyek védelmét biztosító alapelvek, szabályok, eljárások, adatkezelési eszközök és módszerek összessége", [azaz] „az adatvédelem az adatalanyok védelme, az adatbiztonság maguké az adatoké."[16]

Egy, az Accenture által végzett kutatás [17], mely a fogyasztói bizalom mozgatórugóit szándékozta feltárni, kiderítette, hogy a cégek merőben másképpen ítélik meg az adatvédelem fontosságát, mint a fogyasztók. Ezen kívül a megkérdezettek több mint fele már megszakította a kapcsolatát azzal a céggel, amely megítélése szerint nem folytatott megfelelő adatvédelmi politikát.

1.táblázat: A fogyasztók cégek iránti bizalmát leginkább befolyásoló tényezők (Accenture)

	Cégek képviselői	Megkérdezett fogyasztók
Kedvező vevőszolgálati tapasztalatok	43%	26%
A céggel folytatott huzamos kapcsolat	27%	29%
A cég vagy termék jó hírneve	23%	33%
A márka ismerete	6%	3%
Adatvédelmi irányelvek	1%	9%

Az Equifax/Harris [15] egyik kutatása szerint a megkérdezettek 33%-a gondolta úgy 1970-ben, hogy a számítógépes technológia veszélyezteti a személyes adataikat. A több mint húsz évvel később megismételt kutatásban már a megkérdezettek 78%-a vélekedett így.

4.2. Az elemzések alapja: a személyes és a különleges adat

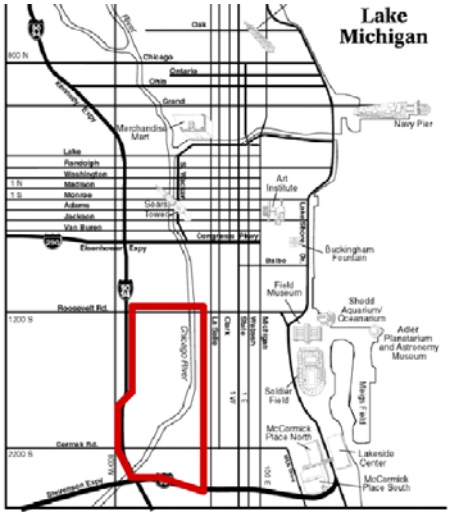
Maga az adat fogalma hiányzik az Adatvédelmi törvényből, az értelmező rendelkezések a személyes adat meghatározásával kezdődnek. Személyes adat a bármely természetes

személlyel kapcsolatba hozható adat, illetve az abból levonható, az érintettre vonatkozó következtetés. Adatbányászat során az elemzések tárgya a személyes adatok tömege, azonban ha a kapcsolatba hozatal már nem valósítható meg, pusztán adatokról beszélünk. Az adat mindaddig személyes adat, míg ez a kapcsolat helyreállítható, az EU irányelv szerint pedig addig, míg ésszerű erőfeszítések árán állítható helyre a kapcsolat. Érdekes, hogy a bírói gyakorlat szerint a távközlési szolgáltatóknál tárolt hívásadat nem számít személyes adatnak. „Meghatározott természetes személlyel kapcsolatba hozható adatnak minősül a természetes személy lakáscíme, telefonszáma. Az az adat azonban, hogy az előfizető telefonvonaláról mikor, mely telefonszám hívásával és milyen időtartamban került sor telefonbeszélgetésre, meghatározott természetes személlyel már nem hozható közvetlenül összefüggésbe. A telefonbeszélgetések időtartama, irányultsága alapján meghatározott természetes személyre következtetés sem vonható le.[1] (BH 2001.269)" [7]

A különleges adatok a fajra, nemzeti, nemzetiségi és etnikai hovatartozásra, politikai véleményre és pártállásra, vallásos vagy világnézeti meggyőződésre vonatkozó adatok, valamint az egészségi állapotra, kóros szenvedélyre, szexuális életre, büntetett előéletre vonatkozó adatok. A különleges adat kezelése az érintett írásos beleegyezése vagy törvényi elrendelés esetén jogszerű (a törvény 2. § 2. *b)* pontjában foglalt adatoknál), illetve akkor, ha az nemzetközi egyezményen vagy alkotmányos jog érvényesítésén alapul, vagy törvény rendeli el a nemzetbiztonság, bűnmegelőzés, illetve bűnüldözés érdekében (a 2. § 2. *a)* pontjában foglalt adatoknál).

Az USA-ban terjedt el a „red lining" kifejezés, mely egy olyan terület határvonala a biztosítócégek és bankok térképein, melyen belül lakó személyek számára nem adható hitel, vagy életbiztosítás. [6] A terület meghatározása a következőképpen történt: megkérték az igénylőket, hogy nyilatkozzanak szomszédságukról, azok faji hovatartozásáról. A módszer érdekessége, hogy a válasz kétfajta információt is szolgáltat, és az igénylőre nézve nem számít különleges adatnak. Először is segít meghúzni a piros vonalat, másrészt a válaszadót is azonosítják ez alapján, hiszen feltételezések szerint az egymáshoz hasonló emberek keresik egymás társaságát, így ha valaki a szomszédját valamely faji kisebbséghez sorolta, rá nézve is hasonlót feltételeztek. Azaz az űrlapon már nem szerepelt a faji eredetet beazonosító rubrika, melyet korábban a cég munkatársa töltött ki, hanem az igénylő nyilatkozott saját szomszédjai faji eredetéről, ami tudta nélkül őt is beazonosította, nyilván nagyobb hibaszázalékkal, mint amikor még az ügyintézők sorolták be „ránézés alapján" az igénylőket. A „red lining" még az után is használatban volt, hogy használatát kifejezetten megtiltották, hiszen azt legtöbbször faji alapon húzták meg, ami diszkriminációt szült.

[1] Megjegyzendő, hogy az adatvédelmi biztos e bírósági jogértelmezéssel teljesen ellentétesen értelmezi ezt a kérdést.



3. ábra: Példa „red lining”-ra [8]

Gyakran találkozhatunk ilyen „trükkös” megkérdezésekkel. Az egyik ilyen gyakori kérdés (az USA-ban) azt firtatja, hogy a megkérdezett személy volt-e már terhes. Ha a válasz igen, abból a nemre is lehet következtetni. Állásinterjú során is gyakran megkérdezik a pályázóktól, hogy öt év múlva hogyan képzelik el a jövőjüket. A válaszból sok minden kiderülhet, női pályázó esetén például az, hogy mikor szándékozik gyermeket vállalni. Egy ártalmatlannak tűnő kérdés sokfelé elvezethet, ami komoly következményekkel járhat.

A faji, nemzetiségi, etnikai alapon történő diszkriminációt egyes források szerint Magyarországon is használták. Automatizált egyedi döntés alkalmazásakor kiküszöbölhető a diszkrimináció, hiszen utólagosan — kérésre — el kell magyarázni a modell működését, meg kell mutatni a modellben szereplő változókat, és a fajra, vagy nemzetiségre vonatkozó változó használatát az intézménynek meg kellene tudni magyaráznia.

Van olyan eset is, amikor az etnikai kisebbségeket pozitívan szeretnék diszkriminálni, és a célpiac megtalálásához szükség lenne adatbányászatra, mely tartalmazná ezeket a különleges adatokat. Ilyen eset volt nem is olyan régen, amikor az egyik németországi távközlési vállalat az országban élő török származású kisebbséget szerette volna megtalálni egy olyan díjcsomaggal, melyben kedvezményes tarifát ajánlottak a Törökországba irányuló hívásokra. Egy kanadai kutatás során pedig arra jöttek rá, hogy a francia ajkú kanadaiak sokkal inkább szeretnek drágább, jobb minőségű benzint vásárolni, mint az angol származásúak. Erre az összefüggésre is szívesen építettek volna, ha a változót felhasználhatták volna. [15] Az ilyen változókból születő modellek azonban a különböző sztereotípiákat tovább erősítenék.

5. OECD irányelvek és az adatbányászat[2]

A Gazdasági Együttműködési és Fejlesztési Szervezet (OECD) Tanácsa alkotta meg az első átfogó adatvédelmi jogi dokumentumot 1980-ban, amelyben megfogalmazott alapelvek sok ország adatvédelmi törvényébe is beépültek. A következőkben ezek az elvek kerülnek bemutatásra, kitérve az elvek és az adatbányászat viszonyára.

Az adatgyűjtés korlátozásának elve
„A személyes adatok gyűjtését korlátozni kell, és ilyen adatok megszerzése csak törvényes és tisztességes eszközökkel történhet, s ha lehetséges, az adatalany tudtával és beleegyezésével."

Ha adatainkat két különböző vállalatnál kezelik, nem feltétlenül szeretnénk azt, hogy ezeket az adatokat közös adatbázisba rendezve, abból ez idáig rejtett összefüggésekre bukkanjanak az elemzők. Ha két vagy több cég mégis ezt szeretné tenni ahhoz, hogy hatékonyabb üzleti döntéseket hozzanak, mindenképpen tájékoztatni kell erről az érintetteket. Az adatvédelmi biztos gyakorlatában is találkozhatunk ilyen esettel: „Egy konzultációs ügyben a Nationale-Nederlanden Magyarország Biztosító Rt. Vezérigazgatója abban a kérdésben kért vizsgálatot, hogy az ING holdinghoz tartozó vállalatok közös használatú adatbázisának kialakítására vonatkozó elképzeléseik az adatvédelmi törvénynek megfelelnek-e. Az adatvédelmi biztos válaszában kifejtette, hogy az adatbázisban szereplő adatok kezelésére és továbbítására csak az érintettek külön és kifejezett hozzájárulása mellett van lehetőség. A holdinghoz tartozó társaságok azonban a hozzájárulás megtagadása esetén sem zárkózhatnak el a szerződés megkötésétől. A hozzájárulás előfeltétele, hogy az érintetteket tájékoztassák arról, hogy a nyilvántartásból pontosan mely vállalatok és milyen célból szerezhetnek adatokat, illetve hogy az adatokat mennyi ideig tárolják. A fentiekre tekintettel az adatvédelmi biztos javasolta, hogy a hozzájáruló nyilatkozat ne váljon a szerződés részévé, hanem az adatkezelők azt külön lapon fogalmazzák meg. Emellett a tervezett hozzájáruló nyilatkozatban az adatkezelők túlságosan tágan határozták meg a kezelhető személyes adatok körét, hiszen az adatkezelés célhoz kötöttségének elve alapján elegendő lenne, ha a nyilvántartásba az érintettnek csupán a név- és lakcímadatai kerülnének. Amennyiben az elkülönített adatbázis létrejöttéig össze kívánják kapcsolni adatbázisaikat, ennek technikai, szervezése és felelősségi kérdéseiről az ügyfeleket tájékoztatni kell. Az adatállományok összekapcsolása azonban még így is sok esetben több, ki nem küszöbölhető veszélyt rejt magában. (33/K/1999.)"

[2] A fejezet megírásához felhasznált irodalmak: [18], [19], [20].

Adatbányászat során sokszor szükség van arra, hogy különböző adatforrásokat összekapcsoljunk. Ilyen lehet az az eset, amikor egy egyénre vonatkozó adatok különböző helyeken tárolódnak. Ha nem kapcsoljuk össze az adatbázisokat, megfosztjuk magunkat a kereszt-kapcsolatok elemzésének lehetőségétől. Összekapcsolás esetén pedig az esetleges duplikációkat kell figyelni. Ezenkívül egy adott helyen tárolt ügyféladat általában egy homogén sokaságból származik, így fontos földrajzi vagy demográfiai megkülönböztetéseket nem tehetünk, csak miután például országos szintű adatokká az előbbieket össze nem fűztük. Gyakorlati példa erre a járványok kitörésének előrejelzése, amikor tipikusan idősoros területi adatokra van szükség. Vállalati gyakorlatban is sokszor találkozhatunk ezzel az igénnyel. A Ford és a Firestone cégek közösen gyártott terméke (Ford Explorer Firestone kerekekkel) esetén például a vállalati titkok és szerződések miatt nem lehetett összekapcsolni a két vállalat adatbázisát, így olyan módszert kellett találni, ami által információhoz lehet jutni, anélkül, hogy a két cég hozzáférne a másik érzékeny adataihoz. [21] Erre fejlesztették ki a későbbiekben tárgyalt PPDM (privacy preserving data mining) módszereket.

Az adatminőség elve
„A személyes adatoknak felhasználási céljaikkal összhangban, és ezeknek a céloknak megfelelő mértében pontosnak, teljesnek és aktuálisnak kell lenniük.”

Minden adatelemzés során szem előtt kell tartani, hogy csak akkor kaphatunk pontos képet, összefüggést, modellt, ha az adataink — amelyeken az elemzésünk nyugszik — jó minőségűek. Ezért nagyon fontos feladat a tudásfeltárás folyamatában az adattisztítás szerepe. Ekkor különböző, általában egyszerű módszerekkel próbálják meg felderíteni, hogy nincs-e hibás adat az elemezni kívánt adatbázisban. Egyszerű gyakoriság kérésével például ellenőrizni lehet egy változó értékeit, kiugró értékek után kutatva. Például egy adott honlapot látogató internetezők nemi megoszlását vizsgálva azt az eredményt kaphatjuk, hogy a látogatók 85%-a nőnemű. Ha női témájú oldalról van szó, az eredmény nem meglepő. Azonban, ha nem így van, rögzítési hibára kell gyanakodni. És valóban, kiderülhet, hogy az űrlapon default értékként a női nem volt beállítva, így annak átállításával valószínűleg sok férfi nem törődött, nem tulajdonítva fontosságot a kérdésnek. [23]

Hand [22] mikrohibáknak nevezi a hiányosan vagy rosszul rögzített adatokat. A makrohibákkal ezzel szemben például a baleseti statisztikák esetében találkozhatunk: míg a nagy, tragikus balesetek körülményeit minden részletre kiterjedően rögzítik, addig a kis balesetekről sokszor nagyon kevés adattal rendelkezünk, vagy azok egyáltalán nem kerülnek rögzítésre. Márpedig ekkor téves következtetésekre juthatunk elemzéseink során.

Az érdekelteknek meg kell adni a lehetőséget arra, hogy téves adataikat korrigálhassák. Az Amerikai Fogyasztók Egyesületének (Consumer Federation of America) már korábban idézett felmérésében [13] különösen sok olyan esetet ismertetnek, melyben rossz adatokat tároltak a bankok ügyfeleikről. A felmérés azt tanácsolja az ügyfeleknek, hogy kb. fél évente ellenőrizzék adataikat, de hitelkérelem intézése előtt mindenképpen. Hibás adat a régi, már nem aktuális adat is, ezeket lassan változó attribútumoknak nevezzük. [24] A hiányos adatok kezelése is kényes kérdés.

A célhozkötöttség elve

„A személyes adatok gyűjtésének célját az adatgyűjtés időpontjánál nem később meg kell adni, és ezt követő felhasználásukat korlátozni kell, mely korlát ezeknek a céloknak vagy ezekkel a célokkal nem összeegyeztethetetlen más céloknak a megvalósulásáig terjedhet, feltéve, ha a megváltoztatott cél minden egyes változás alkalmával meg van adva."

Az elv lényege, hogy az adatkezelésnek előre meghatározott célja legyen és az érintettet tájékoztassák arról, hogy milyen célból gyűjtik, kezelik a rá vonatkozó adatokat. Adatbányászat esetén a célt nehéz meghatározni, hiszen adatbányászat során pontosan a rejtett összefüggések feltárása a cél, olyanoké, melyek létezéséről ez idáig nem volt tudomásunk. Sokan azt javasolják, hogy célként az adatbányászatot kellene megadni, de mint tudjuk, ez egy eszköz, semmiképpen sem a cél. Cél lehet például a testreszabott ajánlatok kialakítása, szegmentálás révén, vagy az ügyfelek minél teljesebb megismerése, amit már az érintettek esetleg elutasítanának, pedig az előbbi esetben is erről van szó. Akkor lennének elégedettek, ha a kinyert információk birtokában a cégek olyan ajánlatokkal állnának elő, melyek mindkét fél számára előnyösek. Cavoukian szerint [18] meg kell adni a lehetőséget az ügyfelek számára, hogy válasszanak a következő listából:

1. nem járulok hozzá az adatbányászati tevékenységhez;
2. hozzájárulok az adatbányászati tevékenységhez, de csak belső, céges használatra;
3. hozzájárulok az adatbányászati tevékenységhez, külső és belső használatra egyaránt.

Ezáltal azonban nagyon leszűkítjük a vállalatok mozgásterét, akik védekezésképpen az eddigi adatbányászati tevékenységüket statisztikai tevékenységnek keresztelhetnék, és tovább folytathatnák azt. A kérdés megoldásra vár még Magyarországon, de az EU-ban is.

A hitelinformációs rendszer kiterjesztésére 2002-ben nagyrészt szintén a célhozkötöttség elvének sérülése miatt nem került sor. „Ebben az évben az adatvédelmi biztos jelentős

eredményeket ért el, melyek közül talán a legfontosabb az, hogy fellépésének köszönhetően a már meglévő negatív listás lakossági hitelinformációs rendszer kibővítésére tett kísérlet nem járt sikerrel. A Pénzügyminisztérium, az Igazságügyi Minisztérium, a Magyar Bankszövetség, és a Magyar Nemzeti Bank, valamint a felügyelet vezetőivel folytatott szakmai konzultáción az adatvédelmi biztos határozottan ellenezte a pozitív listás adósnyilvántartási rendszer felállítását. Ennek értelmében nem csak a minimálbért meghaladó 90 napon túli késedelmes tartozásokról vezetne a bankszektor központi nyilvántartást, hanem minden egyes hitelszerződés teljes adattartalma az adósnyilvántartásba kerülne. Hangsúlyozta, hogy a javasolt szabályozás a személyes adatok védelméhez való jog korlátozásának feltételeit — szükségesség, arányosság, a cél elérésére való alkalmasság — nem elégíti ki. A központi hitelinformációs rendszert legfeljebb azok adataival indokolt kibővíteni, akik korábbi hitelkérelmük kapcsán hamis adatokat szolgáltattak, vagy más hasonló visszaéléseket követtek el, de az ő esetükben is megfelelő adatvédelmi garanciákra lenne szükség. A biztos hivatkozott a francia adatvédelmi biztos 2000. évi beszámolójára is, mely kétségessé teszi a pozitív listás rendszer hatékonyságát, és felhívja arra a veszélyre a figyelmet, hogy a pozitív lista nagyobb teret enged a célhozkötöttségi elv megsértésének, mivel a benne lévő információk nagy száma és gazdagsága komoly kísértés a más célra történő felhasználásra."

Ha a cél megszűnt, törölni kell az adott célból tárolt adatokat is, vagy anonimizálni kell őket. Ennek egyik oka, hogy ha az adatok már nincsenek fókuszban, akkor nagyobb az esély arra, hogy illetéktelenek hozzáférnek, lemásolják, vagy ellopják azt.

A korlátozott felhasználás elve
„A személyes adatokat nem szabad felfedni, hozzáférhetővé tenni, vagy egyéb olyan módon felhasználni, amely eltér a 9. paragrafusban rögzített céloktól, kivéve:
(a) az adatalany beleegyezésével; vagy
(b) ha a törvény azt úgy rendeli."

Adatbányászat során lehetőség van arra, hogy egy bizonyos célból gyűjtött adatot felhasználjunk más célra is. A korlátozott felhasználás elve még a '70-es években született, amikor a többszörös adat-felhasználásra még nem volt lehetőség, vagy igény. Ebben a tekintetben eljárt az idő ezen elv felett, hiszen technikailag minden lehetőség megvan arra, hogy több célból is megvizsgáljuk ugyanazon adatokat.[3] Bankkártya használatánál ha az elsődleges cél az, hogy a pénzfelvétel jóváhagyásra kerüljön, akkor a tranzakciós adatokat

[3] Technikailag valóban megvan a lehetőség (vagy inkább veszély), hogy az egyik célból kezelt adatokat egy másik célra is felhasználjanak — ez azonban korántsem teszi idejétmúlttá azt az alapelvet, hogy személyes adatokat csak az adatalany hozzájárulásával vagy törvényi felhatalmazással lehet kezelni bármilyen célra. *(A szerkesztő megjegyzése.)*

nem lehet arra használni, hogy rajtuk adatbányászatot végezzenek. E másodlagos, jövőbeli célhoz is kérni kell az érintettek hozzájárulását. És mint korábban is tárgyalásra került, az adatbányászat nem cél, hanem eszköz, ami további kérdéseket implikál.

A biztonság elve

„A személyes adatokat ésszerű biztonsági intézkedésekkel védelmezni kell olyan veszélyek ellen, mint elvesztés vagy illetéktelen hozzáférés, megsemmisülés, felhasználás, módosítás vagy megismerés."

A biztonság elvének betartása minden adatkezelésnél fontos feladat, amit adatbányászat során is szem előtt kell tartani.

A nyíltság elve

„A személyes adatokra vonatkozó fejleményeket, a velük folytatott gyakorlatot és politikát nyíltan kell kezelni. A személyes adatok létezésének természetének és felhasználásuk fő céljának, valamint az adatkezelő személyének és állandó tartózkodási helyének megismerésére egyszerű módszereket kell kidolgozni."

A nyíltság elve az egyik legfontosabb alapelv. Az embereknek joguk van ahhoz, hogy megtudják: ki, hol, milyen jogalappal, milyen személyes adatokat kezel. Ezen túlmenően pedig (részben a következő alapelv alapján) az érintetteknek ahhoz is joguk van, hogy megtudják, milyen adatokat tárolnak róluk, kik férhetnek hozzá adataikhoz, és hogyan használják fel azokat. Sok esetben nem is gondolnánk, hogy például videofilm kölcsönzésekor milyen sokat elárulhatunk magunkról. A kölcsönző korábbi kölcsönzéseink alapján megrajzolhatja profilunkat adatbányászat segítségével. Ez amellett, hogy elsőre talán ijesztően hathat, számunkra is előnyökkel járhat. Ahelyett, hogy számunkra érdektelen ajánlatokkal bombáznának bennünket a jövőben, pontosan olyan filmeket fognak számunkra kínálni, amelyeket nagy valószínűség szerint szeretni fogunk, olyan akciókat kínálnak majd, mellyel szimpatizálunk (például hármat vihet, kettőt fizet), és olyan csatornán keresnek majd, melyet a leginkább kedvelünk (e-mail postai út helyett). Cross-sell ajánlatokat is érdeklődésünknek, ízlésünknek megfelelően kínálnak majd.

Az adatbányász munka során általában rengeteg új változót hozhatunk létre. Ezenkívül a modellek lefutásának eredményeként szintén újak születhetnek. Ilyen például a klaszterezés során létrejövő csoportbesorolásokat rögzítő változó. A klaszterezési eljárások során — amelyek éppúgy megtalálók a többváltozós statisztika eszköztárában, mint az adatbányászatéban — az egymáshoz közel álló itemeket (például fogyasztókat, felhasználókat, látogatókat) soroljuk csoportokba úgy, hogy azok a vizsgált változók alapján a leginkább hasonlítsanak egymásra, és a leginkább térjenek el a többi csoporttól.

Általában arra törekszünk, hogy kezelhető darabszámú klaszter szülessen. Az adatbázisba visszaírásra kerül ezután a klasztertagságot jelző változó. Ha egy ilyen adatot tárol egy vállalat ügyfeleiről, kérdés, hogy erről is „el kell-e számolnia".

A személyes részvétel elve
„Az egyénnek legyen joga arra,
(a) hogy az adatkezelőtől vagy másképpen bizonyságot nyerjen arról, van-e az adatkezelőnek rá vonatkozó adata vagy nincs;
(b) hogy a rá vonatkozó adatokat megkapja éspedig
a. elfogadható időn belül;
b. méltányos díj ellenében, ha ilyen egyáltalán van;
c. elfogadható módon; és
d. számára könnyen elérhető formában;
(c) hogy az (a) és (b) alpontok szerinti igényének elutasítása indokait megismerje, és az elutasítását kifogásolja;
(d) hogy kifogásolja a rá vonatkozó adatokat és, ha a kifogásolás helyénvaló, az adatokat töröltesse, helyesbíttesse, kiegészíttesse vagy módosíttassa."

A személyes részvétel elve biztosítja egyrészt a kontrollt, másrészt az adatminőség elvének érvényesülését is segíti. Amíg az érintettek nem tudják, hogy adatbányászat segítségével róluk milyen pontos képet lehet megrajzolni, aminek többé-kevésbé komoly következményei lehetnek, addig nem is nagyon gyakorolják a kontrollt. Téves adatok miatt esetleg valakinek folyamatosan krimiket ajánl a rendszer a videotékában, míg ő csakis romantikus vígjátékokra vágyik. Ez legfeljebb bosszantó tud lenni egy idő után. Ellenben ha valakinek a hitelkérelmét azért utasítják el, mert egy helyiértékkel „elírták" az éves családi jövedelmét, annak már komoly következményei lehetnek. Ezért is fontos, hogy felhívják az érintettek figyelmét arra, hogy mire jó az adatbányászat, és egy adott cégnél mire használják azt, milyen kihatása lehet az érintett életére.

A felelősség elve
„Az adatkezelőnek felelősnek kell lennie azoknak az intézkedéseknek a betartásáért, amelyek a fenti elveket tükrözik."

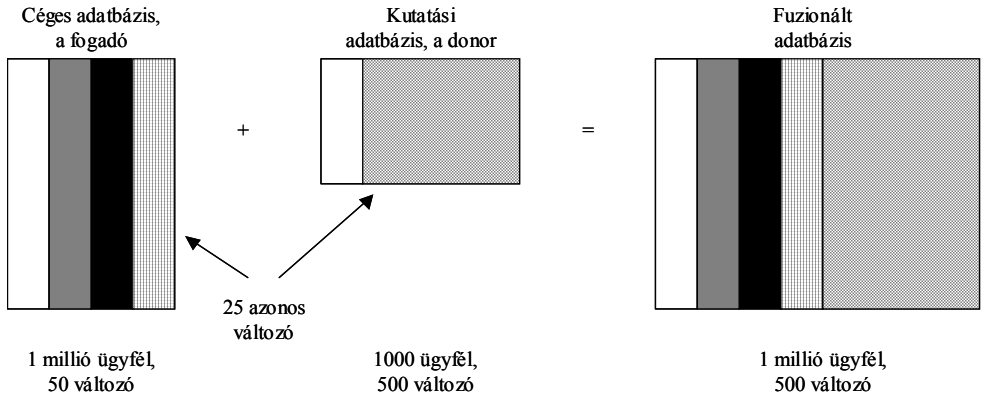
A záró elv felelőssé teszi az adatkezelőt a fenti elvek maradéktalan betartásáért, betartatásáért.

A magyar Adatvédelmi törvény szintén tartalmazza a fenti elveket, a célhozkötöttség elvét, az adatminőség elvét, az adatbiztonság és a személyes részvétel elvét külön pontban tárgyalva.

6. PPDM (Privacy preserving data mining), azaz a személyes adatok védelmét biztosító adatbányászat

Adatvédelemmel kapcsolatos törekvések már az adatbázisok létrehozásakor, karbantartásakor is felmerültek, de akkor kezdtek igazán hangsúlyossá válni, amikor szükségessé vált a különböző helyeken fellelhető adatbázisok összekapcsolása — akár vállalaton belül, akár két vagy több cég között.

Kutatási eredmények szerint [18] az érintettek sokkal inkább tartanak adataiknak külső másodlagos használatától, mint az adott cégen belüli másodlagos használattól. Az adatvédelmi törvény szerint az adattovábbítás és adatkezelés összekapcsolása csak az érintett beleegyezésével történhet meg, vagy ha azt a törvény megengedi. Cégen belül például ilyen összekapcsolásról van szó, amikor a piackutatási adatokat „kivetítjük" a teljes ügyfélbázisunkra. Tegyük fel, hogy rendelkezünk egy milliós ügyféladatbázissal, amiben az ügyfelekről 50 változóban tárolunk adatokat. Ezenkívül készült egy piackutatás is, melynek során ezer ügyfelet kérdeztek meg, ötszáz kérdést feltéve (ennyi változónk van). Ha a két adatbázisban van valamennyi (mondjuk 25) azonos változónk, akkor azokat összekapcsolva, kapunk egy adatbázist, mely egymillió ügyfélre tartalmaz ötszáz változót. Az összekapcsolás alapja, hogy olyan értékeket rendelünk a piackutatásba be nem vont ügyfelekhez, amelyek azon megkérdezetteknél fordultak elő, akik a közös változók tekintetében a legjobban hasonlítanak rájuk. (Ez a vertikális összekapcsolás. Amikor például további ügyfelekkel bővítjük adatbázisunkat, horizontális összekapcsolásról beszélünk, akkor nem az adatbázis „szélessége" nő a változókkal, hanem a „hossza", az ügyfelekkel. [25])



4. ábra: Piackutatási adatok integrálása [19]

6.1. Ötletek az adatbázis-statisztikából

Egy korábbi fejezetben már bemutatásra került az anonim architektúra, valamint az aggregált adatok tárolása — mint két triviális módszer a személyes adatok védelmére. Mindenképpen olyan megoldások kívánatosak, amelyek nem annyira a torzítást minimalizálják, sokkal inkább igyekeznek az általános mintázatokat megőrizni. A módszerek jellemzője, hogy egyfajta átváltást fedezhetünk fel a privacy megvalósulása és a mintázatok torzulásának mértéke között, azaz minél inkább teljesül a személyes adatok védelme, annál kevésbé jó mintázathoz juthatunk az adatokban rejlő információk kinyerése által.

Brankovic és Estivill-Castro [15] két csoportba sorolja ezeket a módszereket:

1. a lekérdezések korlátozása (query restriction)
2. zaj hozzáadása (noise addition).

Lekérdezések korlátozásakor előre megadott lekérdezésekre egzakt választ kapunk, ám elutasításra kerülnek azon lekérdezések, melyek veszélyeztetik a privacy-t. Sokak szerint a módszer gyengéje, hogy a tapasztalt felhasználók, elemzők könnyűszerrel kijátszhatják azt; figyelmen kívül hagyhatnak néhány valóban fontos információt, és elmoshatnak bizonyos mintázatokat. Ilyen kijátszási lehetőséget szolgáltat az átfedő lekérdezések (overlapping query) problémája. Tegyük fel, hogy lehetőségünk van egy vállalat adatbázisában az összes dolgozó átlagos fizetésének lekérdezésére. Továbbá tegyük fel, hogy kérünk egy szűkítést, lekérdezzük a céges átlagfizetést, a vezérigazgató kivételével. A két eredményből megkapjuk a vezérigazgató fizetését, ami már személyes adat, és melyhez nem juthattunk volna hozzá közvetlenül, a korlátozások miatt. Ez az egyszerű példa is szemlélteti, mennyire sérülékeny a megoldás. E kijátszástól való félelem annyira komoly lehet, hogy erősen korlátozott szettet bocsáthatnak az elemzők rendelkezésére, olyannyira korlátosat, hogy az a legtöbb esetben nem is használható adatbányászatra… Ezzel kapcsolatban a kérdés az, hogy az adatbányász miért vizsgálná meg ezt a szettet, ha az garantáltan nem tartalmaz semmilyen hasznos információt? Ráadásul annak biztosítása, hogy a szett valóban ne tartalmazzon semmilyen mintázatot vagy információt, csak úgy lehetséges, ha előzetesen megkeressük ezeket a mintázatokat, vagy nagyon kis szettet biztosítunk. Végül arra is szükség van, hogy az elemzők ne játszanak össze egymással, vagy hogy senki se férjen hozzá a megengedettnél több adathoz.

A második módszer ezzel szemben hibát ad vagy az alapadatokhoz, vagy a lekérdezések eredményéhez. A technika ugyan robosztus, azonban mégis statisztikailag kevésbé

megbízható adatokhoz juthatunk. Ennek iparágtól függően jelentős hatása is lehet: míg sok helyen az 5%-os szignifikancia szint használata teljességgel elfogadott, addig például a gyógyszeriparban — érthető módon — ennél jóval kisebb értékeket követelnek meg.

Brankovic és Estivill-Castro mellett Srikant [29] is említést tesz a data swapping módszerről, ami nem más, mint az adatbázisbeli sorok értékeinek egymással való felcserelése úgy, hogy az alapstatisztikák még megbízható eredményekkel szolgáljanak. Alapja, hogy az egymáshoz legközelebb álló elemek érzékeny adatait felcserélik egymással, így bár hozzáférnek az elemzők a személyes adatokhoz, azok valójában nem az adott egyénhez tartoznak, a legtöbb, amit elmondhatunk róluk, hogy „a valós értékek nagyjából a megjelenített értékek környezetében találhatók". Brankovic és Estivill-Castro sikerrel használta ezt a módszert döntési fák esetében. Azok a változók, amelyek nem érintettek a data swapping által, ugyanúgy helyezkednek el a döntési fán, mint data swapping nélkül. Minél inkább a fa gyökerénél végezzük el a swapping módszerét, annál inkább garantáljuk az adatok védelmét, bár a precizitás mértéke kisebb.

Srikant triviális módszerként említi még az adatbázis mintával való helyettesítését.

6.2. A randomizált pertubáció

Agrawal és Srikant ([25], [27]) azt javasolta, hogy mielőtt az adatokat az elemzők megkapják, gondoskodni kell arról, hogy azokhoz zajt adjanak hozzá. A feladat, hogy olyan metódust kell találni, amely során ennek a zajnak minimális lesz a torzító hatása. [26] Ők ezt a döntési fák használata esetén próbálták megvalósítani, amihez az előbbiekkel ellentétben a randomizált pertubáció módszerét használták.

A módszer lényege a következő:

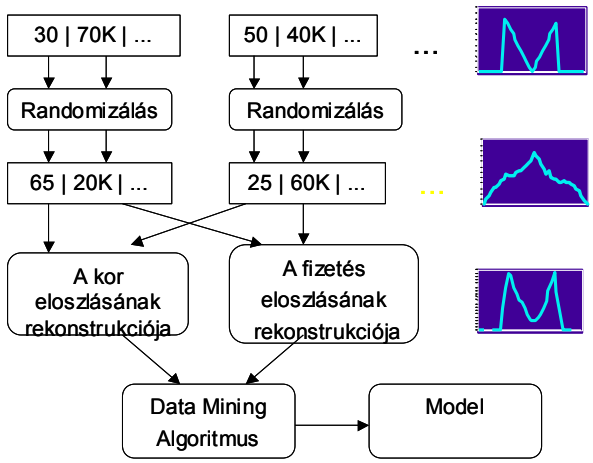
Legyen

x_i: érzékeny adat;

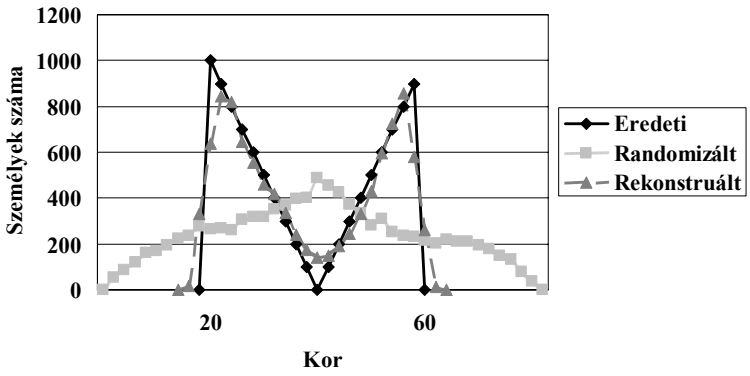
x_i+r: az érzékeny adat adatbázisban megjelenő értéke;

r: véletlen szám.

Tegyük fel, hogy a véletlen szám egy ismert eloszlással rendelkezik (például normális eloszlás). Ismert eloszlás segítségével generálva a véletlen számot, megkaphatjuk az érzékeny adat eloszlását. Az algoritmusok tehát csak megbecsülik, vagy megpróbálják rekonstruálni az eredeti adatszettet, de nem tudják annak pontos értékeit. A változók függetlenségét feltételezve pedig lehetőség van akár döntési fa megalkotására is.



5. ábra: A módszer bemutatása web demográfiai vizsgálaton [29]



6. ábra: A módszer jól működik — Srikant mérése [29]

Du és Zhan [25] új betűszót alkotott: DTPD (Building Decision Tree on Private Data), azaz döntési fa építése személyes adatokon, melyhez szintén a randomizálás technikáját alkalmazták. Ennek menete a következő: „A" adatokat gyűjt egy központi adatbázist alkotva, mely adatbázison adatbányászati tevékenységet szeretne folytatni. Kérdés: „A" hogyan gyűjtsön adatokat úgy, hogy ne tudjon meg túl sokat a megkérdezett személyekről, mégis meglehetősen pontos döntési fát kapjon? Ehhez Du és Zhan a randomizálási technikát választotta, melynek lényege, hogy az adatokat annyira összekeverik, hogy az elemző egy előzetesen definiált valószínűségnél jobban ne tudja megmondani, hogy az

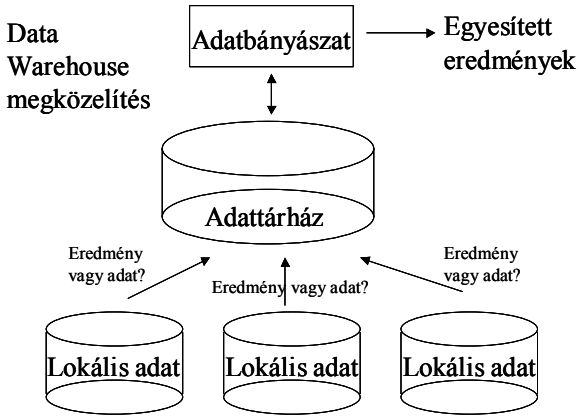
ügyféladat igaz vagy hamis információt hordoz. Ha nagy az elemszám, megfelelő pontossággal nyerhető ki az aggregált információtartalom.

Evfimievski [25] a randomizálás technikáját asszociációs szabályoknál alkalmazta. Nagy eltérés Agravanttól és Srikanttól, hogy minden egyes attribútumot külön-külön, egymástól függetlenül kódoltak, „rejtettek el". Ha az attribútumok száma nagy, az adatminőség szignifikánsan romlik.

6.3. Személyes adatok védelmét biztosító, disztributált adatbányászat (PPDDM)

A PPDM-mel való foglalkozást, kutatást még jobban ösztönözte, amikor felmerült az igény arra, hogy a különböző helyeken fellelhető adatokon adatbányászati műveleteket végezzenek (így került még egy D, a disztributált rövidítéseként a PPDDM-be.) Az alábbi kérdések merülnek fel ezzel kapcsolatban [8]:

- *Adatok összekapcsolása* — Megoldható-e a különböző helyeken tárolt adatok összekapcsolása?
- *Heterogenitás* — Mi az egyszerűbb, az eredmények vagy a források összekapcsolása?
- *Privacy* — A különböző szervezetek vagy szervezeti egységek szívesen megosztják egymással adatbányászati eredményeiket, azonban az adataikat nem. (Példák találhatók az „Adatgyűjtés korlátozásának elve" pont alatt.)



7. ábra: DW megközelítés — eredmény vagy adat? [8]

A továbbiakban ez utóbbi kérdéskörre koncentrálunk: lokálisan elvégezzük az adatbányászatot, majd az eredményeket összesítjük anélkül, hogy az adatokat megosztanánk a források között.

Az asszociációs szabályok[4] esetét vizsgálva a következő lemmát alkothatjuk (bizonyítását lásd a forrásban [8]): ha egy szabály support értéke nagyobb k%-nál globálisan, akkor legalább az egyik lokális helyen a support érték nagyobb k%-nál.

Ez alapján a disztributált algoritmus működése a következő: az összes lokális hely küldje el azon szabályait, ahol a support értéke legalább k%. Ezekhez a szabályokhoz kérjük el a különböző elemszámokat, hogy minden szabály globális support értékét megállapíthassuk. Ekkor biztosak lehetünk abban, hogy minden szabályt megtaláltunk, melynek support értéke nagyobb k%-nál. Ez a módszer mindaddig működik, míg az érdekelt felek hajlandóak megosztani egymással a support értékeket.

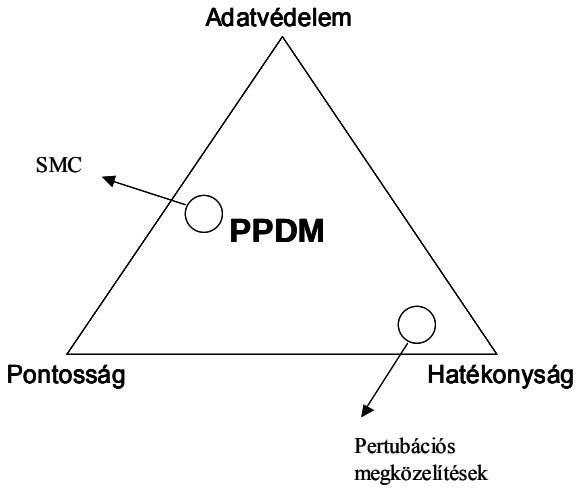
2000-ben alkalmazta Lindell és Pinkas [26] először az SMC, azaz a Secure Multi-party Computation módszerét döntési fákon. Ilyenkor két vagy több lokális helyen vannak az adatok, mely helyek egymással kooperálnak, hogy hozzájussanak a globális adatbányászati eredményekhez anélkül, hogy felfednék adataikat. Lényege, hogy a számítások végén egyik fél se tudjon többet az adatokról, mint a saját inputja és a létrehozott modell. Ennek egy lehetséges módját mutatják be az asszociációs szabályokkal kapcsolatban leírtak; egy másik lehetőség, ha létezik egy olyan fél, akiben a többiek megbíznak, és aki elvégzi a számításokat, a cégek adatbázisainak összekapcsolását, és a modell létrehozása után megsemmisíti a közös adatbázist, a modellt pedig minden érintett fél tudomására hozza. Ezt a módszert megbízott harmadik feles modellnek nevezte el a szakirodalom. [25] Ez utóbbi módszer nehézsége, hogy nem könnyű olyan harmadik felet találni, akiben meg lehet bízni, hogy nem adja ki a bizalmas személyes adatokat, illetve el is tudja végezni az elemzést professzionális szinten. A később született munkák zöme azzal foglalkozik, hogy valamilyen úton kiválthassák ezt a harmadik felet (például lásd [30]), többek közt Lindell és Pinkas tanulmánya is. Ennél a nem központosított megoldásnál jelentős performancia problémák is felléphetnek a lokális források számának növekedésével.

[4] „Legyen τ az ι hatványhalmaza felett értelmezett sorozat. Az R: $I_1 \rightarrow I_2$ kifejezést a c bizonyosságú, s támogatottságú asszociációs szabálynak nevezzük, ha I_1, I_2 diszjunkt elemhalmazok, és $c = supp_\tau (I_1 \cup I_2) / supp_\tau (I_1)$, $s = supp_\tau (I_1 \cup I_2)$. A szabály bal oldalát feltétel résznek, jobb oldalát pedig következmény résznek nevezzük. Az R: $I_1 \rightarrow I_2$ szabály bizonyosságára gyakran conf (R)-ként hivatkozunk.
Feladat egy adott kosársorozatban azon asszociációs szabályok megtalálása, amelyek gyakoriak (támogatottságuk legalább min_supp), és bizonyosságuk egy előre megadott korlát felett van. Jelöljük ezt a bizonyossági korlátot min_conf-fal." [21]

6.4. Összegzés

A modellek tökéletesítése során rájöttek, hogy az adatvédelem, a pontosság és a hatékonyság csak egymás kárára javítható. A későbbi megoldásokat e három tényező szempontjából minősítették. Minden adatbányászati modellhez kapcsolódóan megpróbáltak olyan módszereket kidolgozni, melyek mindhárom tényezőnek egyaránt megfelelnek. (A módszerek hatékonyságának javításával főleg [31] és [32] foglalkozik.) Ilyen módszereket főleg a döntési fák és az asszociációs szabálygenerálás területein dolgoztak ki. (Erről bővebben lásd [27], [28].)

A következő ábra a PPDM két legfontosabb megoldását helyezi el a tényezők viszonylatában, a pertubációs megközelítéseket, mely Agrawalnál és Srikantnál jelentek meg először és az SMC-t, mely Lindell és Pinkas nevéhez fűződik. Mindkét módszer 2000-ben jelent meg, és mindkettő hozzájárult a PPDM megalapozásához. [26]



8. ábra: Trade-off a tényezők között [26]

Clifton [8] kísérelte meg rendszerezni a PPDM módszereit. Ő három csoportba sorolta a fent bemutatottakat, rengeteg szemléltető példát összegyűjtve:

1. *Adatok megváltoztatása* — véletlenszerű módosítás, értékek felcserélése (data swapping), kontrollált adatmódosítás, hogy a valós adatokat senki se lássa. Probléma, hogy egyrészt nem feltétlenül védettek így az adatok, másrészt

kérdéses az ilyen adatok elemzéséből kapott eredmények használhatósága. (Lásd erről bővebben [15] és [29]).

2. *Aggregálás, összefoglalás* — statisztikák kérése, lekérdezések korlátozása. Itt is kérdéses, hogy milyen adatbányászati műveletet lehet végrehajtani az ilyen aggregált adatokon, elég-e a kapott információ, van-e lehetőség következtetések levonására?
3. *Szeparáció* — megbízott harmadik feles módszer. Valóban megcsinálja a harmadik fél az elemzéseket? Nem él vissza a helyzetével? Kiváltható a harmadik fél?

A rendszerezésből hiányzik néhány korábban említett megoldás, így talán az összegzés szempontjából célszerűbb egy másfajta kategorizálást bevezetni:

1. Triviális módszerek, melyek az adatbázis-statisztikában is megtalálhatók
    a. Anonim architektúra (azonosítók törlése, átkódolása)
    b. Adatbázis helyett minta használata
    c. Aggregált adatok tárolása
    d. Lekérdezések korlátozása (query restriction)
2. Zaj hozzáadása
    a. Kontrolált adatmódosítás
    b. Data swapping
    c. Randomizált pertubáció
        i. Például DTPD, asszociációs szabályokon
3. PPDDM
    a. SMC.

7. Záró megjegyzés

Az adatbányászati tevékenység még nem mindennapos gyakorlat a magyarországi vállalatoknál, használatát csak a legnagyobbak engedhetik meg maguknak. (Mindamellett, hogy a kis- és közepes vállalatok zöme esetében nincs is rá szükség.) Emiatt az adatvédelem és az adatbányászat kapcsolatát még csak kevesen vizsgálják. A nemzetközi szakirodalom már „ontja” a javaslatokat, módszereket, megoldásokat, de azok is kiforratlanok, kevés gyakorlati esetben tesztelték őket. Ha vállalati vagy szabályozási oldalról megnő az igény az adatvédelmet biztosító adatbányászati megoldások iránt, valószínűleg itthon is több kutató fog a témával foglalkozni. Addig is a hagyományos módszerekkel vigyázzák a vállalatok adataink védelmét.

Irodalomjegyzék

[1] Adriaans, P. – Zantinge, D. (2002.): *Adatbányászat.* Panem Könyvkiadó Kft.

[2] Berry, M. J. A. – Linoff, G. (1997): *Data Mining Techniques for Marketing, Sales, and Customer Support.* Wiley Computer Publishing.

[3] Stifán Orsolya (2004.): Adatbányászat és statisztika. In: *Alma Mater. Logisztika, információmenedzsment, szoftvertechnológia.* BME GTK Információ- és Tudásmenedzsment Tanszék.

[4] Hajdú L – Mundruczó Gy. – Vita L. (1997.): *Statisztika.* AULA.

[5] Dr. Oros Paulina – Dr. Szurday Kinga: *Adatvédelem az Európai Unióban. Szolgáltatások szabad áramlása.* Európai Füzetek 35. Szakmai összefoglaló a magyar csatlakozási tárgyalások lezárt fejezeteiből.

[6] Berson, A. – Smith, S. – Thearling, K. (1999): *Building Data Mining Applications for CRM.* McGraw-Hill.

[7] Dr. Jóri András: Az adatvédelemről rendszergazdáknak. http://www.jogiforum.hu/files/publikaciok/jori_adatv_rendgazdaknak(jf).rtf

[8] Clifton, C.: Lay of the land: Legal, Moral, and Historical reasons why Privacy Preserving Data Mining is Important, www.cs.purdue.edu/people/clifton

[9] Szabó, M.: Távközlési forgalmi adatok az Európai Unióban http://gportal.hu/portal/adatvedelem/ 2003.10.31. 16:00

[10] Az adatvédelmi biztos ajánlásai, 2000. (99/A/2000, 106/A/2000, 120/A/2000, 300/A/2000, 404/A/2000, 602/A/2000)

[11] Komrád, K. (2002): On credit scoring estimation. Humboldt University, Berlin.

[12] Pannon GSM Távközlési Rt. Általános Szerződési Feltételek. http://pgsm.hu/ertekesites/preusz040401.pdf

[13] Credit Score Accuracy and Implications for Consumers, December 17, 2002. Consumer Federation of America, National Credit Reporting Association. http://www.consumerfed.org/121702CFA_NCRA_Credit_Score_Report_Final.pdf

[14] 1992. évi LXIII. törvény a személyes adatok védelméről és a közérdekű adatok nyilvánosságáról

[15] Brankovic, L., Estivill-Castro, V.: Privacy Issues in Knowledge Discovery and Data Mining. http://www.sct.gu.edu.au/~s2130677/teaching/KDD.d/readings.d/AICE99.pdf

[16] Székely Iván: Adatvédelem és nyilvánosság. In: *Távközlő hálózatok és informatikai szolgáltatások.* http://www.hte.hu/onlinekonyv.html

[17] Statisztika – a hét statisztikái. Mitől függ a fogyasztói bizalom? http://www.ittk.hu/infinit/2004/0304/indexst2.html

[18] Cavoukian, Ann: Data mining: Staking a Claim on Your Privacy. 1998. www.ipc.on.ca

[19] Marten den Uyl, J. – Holleman P.B. – Rob van der Veer, J.: Impact of Data Mining on Privacy and ISAT. 2001. Privacy Incorporated Software Agent. Internal Report. http://www.pet-pisa.nl/pisa_org/pisa/pisa_project_key_issues.html

[20] Az OECD Tanács ajánlása a magánélet védelmét és a személyes adatok határátlépő áramlását szabályozó irányelvekre. (1980. szeptember 30.)

[21] Bodon, F.: Adatbányászati algoritmusok. (2002–2004.) http://www.cs.bme.hu/~bodon/magyar/adatbanyaszat/tanulmany/adatbanyaszat.pdf

[22] Hand, D. J.: Statistics and Data Mining Intersecting Disciplines. *SIGKDD Explorations,* Volume 1, Issue 1. 1999.

[23] Kohavi, R., Brodley C., Frasca, B., Mason L., Zheng, Z. (2000): KDD-Cup 2000 Organizers' Report: Peeling the Onion. In *SIGKDD Explorations* 2:2, 86–98. ACM Press. http://robotics.stanford.edu/users/ronnyk/ronnyk-bib.html, http://www.lsmason.com/papers/SIGKDD00-KDDCup2000.pdf

[24] Kohavi, R., Mason, L., Parekh, R., Zheng, Z.: Lessons and Challenges from Mining Retail E-Commerce Data. *Machine Learning Journal,* Special Issue on Data Mining Lessons Learned, 2004. http://maya.cs.depaul.edu/~classes/ect584/papers/kohavi.pdf

[25] Du, W. – Zhan, Z.: Building Decision Tree Classifier on Private Data. 2003. http://www.cs.purdue.edu/homes/clifton/icdm02/talks/du.pdf

[26] Vaidya, J – Kantarcioglu, M..: An Architecture for Privacy Preserving Mining of Client Information. http://www.cs.purdue.edu/homes/jsvaidya/pub-papers/icdm02psdm.pdf

[27] Da, W. – Zhan, Z.: Using Randomized Response Techniques for Privacy-Preserving Data Mining. *SIGKDD '03,* August 24–27, 2003, Washington DC, USA.

[28] Oliveira, S.R.M. – Zaiane, O. R.: Algorithms for Balancing Privacy and Knowledge Discovery in Association Rule Mining. http://web.cs.ualberta.ca/~zaiane/postscript/ideas03-1.pdf

[29] Srikant, R.: Privacy Preserving Data Mining: Challenges and Opportunities. http://www.almaden.ibm.com/cs/people/srikant/talks/pakdd.ppt

[30] Clifton, C. – Kantarcioglu, M. – Vaidya, J – Lin, X. – Zhu, M.: Tools for Privacy Preserving Distributed Data Mining. SIGKDD Explorations. Volume 4, Issue 2. http://www.cs.purdue.edu/homes/clifton/DistDM/kddexp.pdf

[31] Wu, C. W.: Privacy Preserving Data Mining: a Signal Processing Perspective and a Simple Data Pertubation Protocol. IBM Research Division. http://www.cis.syr.edu/~wedu/ppdm2003/papers/2.pdf

[32] Agrawal, S. – Krishnan, V. – Haritsa, J. R.: On Addressing Efficiency Concerns in Privacy-Preserving Mining. http://arxiv.org/abs/cs/0310038