Egészségügyi adatok előkészítése elemzések céljából

Dr. Fogarassyné Vathy Ágnes, Veszprémi Egyetem
Dr. Fogarassy György, Állami Szívkórház, Balatonfüred

Az orvostudomány mindig törekedett arra, hogy gyógyító tevékenységét tényekre alapozza. E tekintetben áttörést jelentett a nagy beteglétszámra, vagy a populáció egy szegmensének egészére kiterjedő megfigyeléses vizsgálatok megindulása. Az utóbbi legkiemelkedőbb példája a Framingham tanulmány, melynek során egy amerikai kisváros lakosságát tanulmányozták egészségi állapotuk, illetve megbetegedéseik szempontjából. Az összegyűlt adatok elemzése révén a 80-as évek óta számos eredményt publikáltak. Ezen vizsgálat során rögzített adatok feldolgozásának segítségével azonosították például a fejlett társadalmakat pusztító koszorúsér eredetű szívbetegség rizikófaktorait is. Azóta az ilyen jellegű adatgyűjtések, illetve elemzések egyre nagyobb intenzitással folynak világszerte (pl. [6]).

Napjainkban számos olyan gyógyszer-alkalmazási vizsgálat zajlik, amelyekben a készítmény adagolásából leginkább profitáló betegcsoportokat próbálják azonosítani. Ezekben a tevékenységekben felhasználják az informatikai rendszerek nyújtotta legújabb módszereket, beleértve az adatbányászati módszereket is, mint például összefüggés-függetlenség analízis, alcsoport-analízis stb.

Azonban új ismereteket nemcsak ilyen vizsgálatok szervezésével nyerhetünk. Mára az egészségügyet már a napi munka szintjén behálózza az informatika, szinte nélkülözhetetlen részévé váltak a napi operatív feladatokat ellátó adatbázisrendszerek. Használatuk során ezek a rendszerek – illetve az archiválásuk révén keletkező adathordozó eszközök – több évre, évtizedre visszamenőleg nagy mennyiségű adatot tárolnak a betegek fizikális állapotáról, életvezetéséről, megbetegedéseiről és gyógyszeres, illetve non-farmakológiai kezeléseiről, halálozásáról. Ezek az információhalmazok olykor csak hasztalan adathalmazoknak tűnhetnek, holott nem tudhatjuk, hogy vajon nem rejtenek-e új, eddig még fel nem tárt összefüggéseket. Ezen adatbázisok vélhetőleg óriási kincsesbányaként szolgálhatnak az orvosi kutatások szempontjából. Általuk igazolhatunk már ismert összefüggéseket, felfedezhetünk kivételes eseteket, illetve segítségükkel akár új összefüggésekre is fény derülhet.

Az elektronikus feldolgozás révén ezen ismeretek feltárása napjainkra már elérhető közelségbe került. Az adatokat most már nem súlyos tömegű papírhalmokból kell összegyűjtenünk, s „véletlenszerűen” meglátnunk az összefüggéseket, hanem megfelelő adatfeldolgozó rutinok és tárolási formák segítségével rábízhatjuk ezt a feladatot a számítógépre. Ehhez nyújtanak megfelelő eszközt az adattárház (OLAP – On Line Analitical Process) és adatbányászó (DM – Data Mining) technikákat implementáló szoftvereszközök. Mindehhez azonban az adatokat össze kell gyűjteni, meg kell őket tisztítani, s a célnak megfelelő strukturális és tartalmi átalakításokat kell rajtuk végezni.

A kórházi információs rendszerek adatainak adatbányászati módszerekkel történő elemzésére ma már sok példát találunk a szakirodalmakban, azonban ezen kutatások nagy része sajnos még csak külföldön folyik (pl. [4], [6]).

Jelen cikk elsődleges célja az egészségügyi adatbázisok elemzésre történő előkészítő tevékenységeinek áttekintése, a felmerülő problémák ismertetése.

AZ ADATELŐKÉSZÍTÉSI FOLYAMAT ÁTTEKINTÉSE

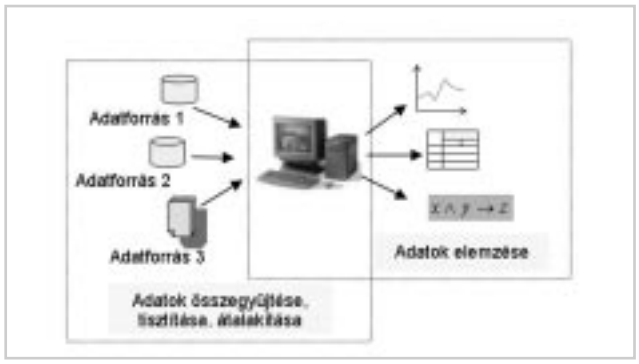
Mint számos más országra, Magyarországra is jellemző, hogy az egészségügyi intézmények gyakran eltérő adatbázisrendszereket használnak adataik tárolására és kezelésére. Ezek a rendszerek azonban nem csak felhasználói felületeikben és „üzleti” logikájukban különböznek egymástól, hanem az egységes egészségügyi információs adatmodell hiányában az adatok tárolási struktúrája sem egységes. Továbbá igaz az is, hogy ezen rendszerek egymástól szeparáltan működnek, tehát köztük az információcsere nem megoldott. Márpedig ha a betegek teljes kórtörténetét össze szeretnénk gyűjteni, s ez által megbetegedéseiket nyomon követni, akkor ezen információs rendszerek adatait egységes adatbázisrendszerekbe, illetve adattárházakba kell egyesítenünk. Hasonló probléma előtt állunk akkor is, ha adatgyűjtésünket csupán egy szakterületre korlátozzuk.

Mivel az adatok rögzítése az esetek nagy részében emberi beavatkozás által történik, így sohasem lehetünk biztosak abban, hogy rendszereink minden esetben valós adatokat tárolnak. Ezért indokoltnak tűnik az adatok tisztításának elvégzése. Továbbá mindezek mellett gondolnunk kell arra is, hogy, az adatok tárolási formája elsősorban a mindennapi tranzakciós tevékenységek elvégzésére lett optimalizálva, ezért az adatelemző és adatbányászó tevékenységek gyakran igénylik az adatok strukturális és tartalmi átalakítását is.

Láthatjuk tehát, hogy az adattárház rendszerek által biztosított elemző tevékenységek és az adatbányászati ismeretfeltárás folyamata a következő adatelőkészítési tevékenységeket igénylik:

- az adatok összegyűjtése és migrálása
- az adatok tisztítása
- az adatok átalakítása az elemzések céljának megfelelően

Az adatelőkészítést követően már következhet az adatok elemzésének és az eredmények értékelésének fázisa. A teljes folyamatot az alábbi ábra szemlélteti.



1. ábra
Az adatok előkészítésének és elemzésének kapcsolata

Az első és a második lépésben megjelölt tevékenységek (adatgyűjtés és migráció, valamint adattisztítás) nem feltétlenül jelentik az előkészítési folyamat sorrendiségét. A sorrendet nagymértékben befolyásolja, hogy a migráció alapját képező adatok azonos struktúrából származnak-e, illetve különbözőekből. Abban az esetben ugyanis ha a forrásadatok struktúrája nagy mértékben hasonló (például ha az adatok ugyanolyan kórházi információs rendszerekből származnak), akkor időtakarékosabb megoldásnak tűnhet először végrehajtani a migrációt, vagyis a több lokális adatbázis adatait egy adatbázisba egyesíteni, majd az adattisztítást ezen az egy, nagy mennyiségű adatot tartalmazó adatbázison elvégezni. Abban az esetben viszont, ha a forrás adatbázisok adatszerkezetei nagymértékben eltérnek egymástól, elképzelhető, hogy először az egyes különálló adatbázisokban végezzük el az adatok tisztítását, majd a migrációt már csak a megtisztított adatokon hajtjuk végre. Továbbá, függetlenül az adatszerkezettől, ha a rendelkezésünkre álló adatok jelentős része szennyezett, szintén ajánlatos a migráció előtt elvégezni az adattisztítást. Így ugyanis a tisztítás után rendelkezésünkre álló adatok már homogénebbek, vagyis a korábban még meglévő szennyeződések már nem befolyásolják a migráció eredményességét (pl. adattranszformációs problémák, mezőméret korlátozások). Azokban az esetekben tehát, ha a szennyeződések száma nagy, és típusai nem előreláthatók, illetve a forrásadatbázisok struktúrája eltérő szerkezetű, célszerű először elvégezni a tisztítást, s majd csak ezt követően migrálni az adatokat. Természetesen gyakran előfordul, hogy az adattisztítás folyamata részben a migráció előtt, részben a migráció után kerül végrehajtásra.

Tekintsük most át a következőkben részletesebben a fenti három tevékenységet.

AZ ADATOK ÖSSZEGYŰJTÉSE ÉS MIGRÁLÁSA

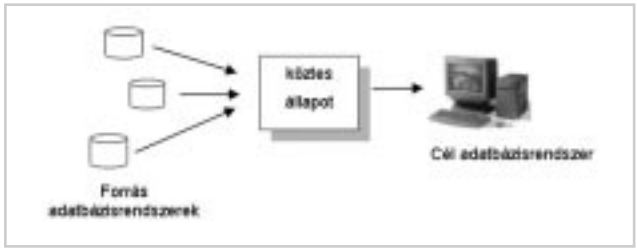
Az adatok migrációja során az adatelemzés céljából a különálló adatokat egyetlen adatbázisba gyűjtjük össze. Ezen adatbázis típusa általában relációs, vagyis az adataink logikailag továbbra is táblázatos formában érhetők el.

Amennyiben az adatok elemzése csupán egyetlen egészségügyi intézmény adatainak elemzésére terjed ki, akkor a migráció lépése akár ki is hagyható, feltételezve, hogy az adatok egyetlen rendszerben állnak rendelkezésünkre. Amennyiben viszont az adataink különböző adatbázisokból érhetők el, az adatok migrációja elengedhetetlen. Az egészségügyi rendszerek migrációja során elmondhatjuk, hogy az esetek jelentős hányadában az adatok különféle kórház-információs rendszerekben, illetve népegészségügyi adatbázisokban állnak rendelkezésünkre. Ettől eltérő, és ezáltal egyszerűbb helyzetbe kerülhetünk akkor, ha például egy intézmény adatait szeretnénk elemezni, illetve ha valamely egészségügyi szolgáltató egységes információs rendszere szolgáltatja az adatokat számunkra.

A migráció folyamata a következő két fő lépésre tagolgató:

- Forrásrendszerek adatainak feltérképezése
- Az adatok transzformálása az új rendszerbe

A fenti két lépés közé gyakran beiktatódik egy közbeeső fázis is, amely során átmenetileg egy köztes területre gyűjtjük össze az adatainkat. A teljes folyamat tehát a következőképpen ábrázolható:



2. ábra
Adatmigráció köztes állapottal

Az adatok migrálásának teljes folyamata részletesebben kifejtve az alábbi lépéseket foglalja magában:

- A forrásrendszerek típusának és adatstruktúrájának meghatározása, melynek során fel kell térképeznünk, hogy
  - a rendelkezésünkre álló adatokat milyen rendszerekben tárolják (pl. Oracle, Ms Access, DB2, Sybase),
  - milyen adatokat tárolnak az egyes rendszerek,
  - továbbá, hogy az adatokat milyen metaadatokkal (pl. adattípus, értékkorlátozások, adatformátum) lehet leírni.
- A forrásadatok közös metszetének megtalálása és a migrálandó adatok körének meghatározása, mely magában foglalja
  - a forrásrendszer mezőelnevezéseiben fellelhető homonímiák és szinonímiák feloldását,
  - a forrásrendszerek adatmezőinek egymásnak való megfeleltetését, valamint
  - az elemzés szempontjából fontos adatok körének meghatározását, beleértve újabb adatbázisok bevonását is, amennyiben szükséges.
- A céladatbázis struktúrájának megtervezése az előző lépés alapján, és az adatszerkezetek tényleges kialakítása.

- A forrásadatok és a cél adatbázisrendszer struktúrájának megfeleltetése, amely a következő fő tevékenységeket foglalja magában:
  - homonímiák és szinonímiák feloldás a forrás és célrendszerek között,
  - a forrás és célrendszerek adatmezőinek egymáshoz rendelése, szükség esetén, bizonyos adatmezők egyesítése, illetve szétbontása,
  - a forrásrendszerek adatainak ellenőrzése, hogy értékeikben megfelelnek-e a célrendszer adattípusának, és amennyiben szükséges az adatok konvertálása, kódolása.
- A forrásrendszerek adatainak kinyerése a köztes adattárolási területre.
- Az adatok importálása a köztes adattárolási területről a célrendszerbe.

Amennyiben szükséges az adatok migrálása, akkor a forrásrendszerek típusának meghatározása és a rendelkezésünkre álló adatok feltérképezése után célszerű az adatokat csoportosítani abból a célból, hogy a különféle rendszerekben tárolt adatoknak meghatározhassuk a közös metszetét. Ez megkönnyíti a szinonímiák és homonímiák feloldását, illetve az egyes adatmezők egymásnak való megfeleltetését. Az egészségügyi rendszereket tekintve ezek a csoportok például a következők lehetnek:

- Betegek személyi adatai (név, lakcím, TAJ szám, születési dátum, …)
- Fizikális státusz (testsúly, testmagasság, …)
- Személyes anamnézis (előző betegségek)
- Rizikó státusz (dohányzás, alkoholfogyasztás, családi anamnézis, …)
- Jelen panaszok
- Diagnózisok (felvételt indokló fődiagnózis, záródiagnózis)
- Beavatkozások (beavatkozás időpontja, típusa, beavatkozást végző orvos, …)
- Vizsgálatok (laboratóriumi adatok, képalkotó vizsgálatok eredményei, …)
- Epikrízis
- Egészségügyi dolgozók adatai (név, beosztás, elérhetőség, …)
- Egészségügyi szolgáltatók adatai (szolgáltató neve, ÁNTSZ kódja, szolgáltató címe, osztály vagy részleg belső kódja, …)
- Egyéb törzsadatok (országok, települések, …)

A rendelkezésünkre álló adatok csoportosítása után szükséges megvizsgálnunk, hogy melyek azok az adatok, amelyek fontosak lehetnek az elemzések szempontjából, vagyis átvitelre kerülnek a célrendszerbe.

A migrálandó adatok kijelölését követően dönteni kell a céladatbázis struktúrájáról. Az adatok közti logikai kapcsolatok feltérképezésével meghatározhatjuk a kialakítandó adattáblákat, az adatmezőket és az adattáblák között lévő kapcsolatokat. Az adatmezők fizikai tervezésénél figyelembe kell venni a rendelkezésünkre álló adatok értékeit is. A kialakítandó mezőket minden esetben maximális méretűre kell tervezni azért, hogy egyetlen értékes adatot se vesszünk el az adatok importálása során. Továbbá a migráció során figyelembe kell vennünk az adatok típusát is. Amennyiben ugyanazt a jelentést hordozó adataink a különböző forrásrendszerekben más és más típust képviselnek két lehetséges megoldás közül választhatunk: a) A céladatbázisban ennek a mezőnek a típusát szöveg típusra állítjuk, s így minden forrásbeli adatot szöveggé konvertálva viszünk át. Ebben az esetben a migrációt követően kell gondoskodnunk az adatok megfelelő átalakításáról. b) A forrásadatbázisban egységes adattípusra hozzuk az azonos adatokat, s így a céladatbázisban minden mező már a végleges adattípusúként kerülhet kialakításra. Az ilyen jellegű adatok konvertálásának szükségszerűségére a tisztítási feladatokat bemutató fejezetben mutatunk példát.

Amennyiben úgy döntünk, hogy adataink a forrásrendszerekben nem igényelnek további átalakításokat, úgy készen állnak a migrációra. A fentiekben már vázoltuk, hogy célszerű az adatokat egy köztes területre „kigyűjteni”, s majd csak ezt követően importálni őket a céladatbázisba. Ezt a köztes lépést általában azért szükséges megtennünk, mivel a cél adatbázisrendszerek nem minden esetben támogatják az adatok bármilyen más rendszerből történő átvételét. Azonban minden adatbázisrendszerre jellemző, hogy támogatja az adatfájlokban tárolt szövegadatok beolvasását. Így hát célszerű adatainkat a forrásrendszerekből először például szövegfájlokba exportálni, majd onnan importálni őket a célrendszerbe. A szövegfájlok formátuma legegyszerűbb esetben vesszővel szeparált (CSV) formátum. (Például: „Varga János, 1963.12.01., 185699975, Budapest Kispipacs u. 1.”) A szeparátorjel szerepét azonban megegyezés alapján nem csak vessző, hanem egyéb karakter is betöltheti. Mivel szöveges adataink gyakran tartalmaznak vesszőt, ezért érdemesebb egy kevésbé gyakran előforduló karaktert választanunk erre a célra, így például a pontosvesszőt. Azonban ebben az esetben is gondoskodnunk kell arról, hogy forrásrendszerek adatértékei ezt a szeparátorjelet ne tartalmazzák. Továbbá, mivel az adatsorok a CSV formátum esetén egy fájlon belül kocsivissza/soremeléssel vannak elválasztva, ezért ez a karaktersorozat sem fordulhat elő a migrálandó adatok értékeiben. Sajnos az egészségügyi adatbázisok adatai gyakran tartalmazzák ezeket a karaktereket, illetve karakterkombinációkat. Gondoljunk csak arra, hogy a gyógyszerérzékenységi adatok tárolása gyakran ölti a következő formátumot: „Lidocain; Novocain; Jód; Tetran; Algopyrin”. Azon túl, hogy az előző adatsor kódolási problémákat is felvet, a CSV fájlban – az adatfelvitel során begépelt „ ;” szeparátorjel miatt – nem egyetlen adatnak, hanem 5 adatnak minősül, amivel egy beolvasó rutin sajnos nem tud mit kezdeni. Hasonló problémát okoz az is, ha megjegyzés típusú mezőben a jobb áttekinthetőség kedvéért az egymástól különálló adatokat új sorba gépelve vitték be a felhasználók.

A migráció utolsó fázisa az adatok beolvasása köztes területről. Ez a lépés bármely megvalósítási formát is választjuk, beolvasási rutinok segítségével már könnyen megoldható.

Térjünk most át a cikk elején ismertetett migrációs folyamat második kérdéskörére, az adatok tisztítására.

AZ ADATOK TISZTÍTÁSA

Az adatok tisztítása lényeges momentum az adatelemzések szempontjából. Természetesen más, s valószínűleg nem valid eredményekhez jutunk elemzéseink során akkor, ha algoritmusainkat, számításainkat eleve rossz adatokon futtatjuk le, végezzük el (GIGO elv – Garbage In, Garbage Out). Gondoljunk csak arra, hogy például milyen nagymértékben befolyásolják már az egyszerű átlagszámítás végeredményét is az egy irányba kiugró értékek. Éppen ezért szükséges, hogy az adatainkat az elemzések megkezdése előtt megtisztítsuk.

Az adattisztítási tevékenységeinket a következő fő csoportokba sorolhatjuk be:

- Numerikus és nem numerikus adatokra vonatkozó értelmezési tartományok ellenőrzése (pl. numerikus adatok minimum és maximum értéke, listaszerű adatok)
- Numerikus adatok ellenőrzése abból a célból, hogy az adatok azonos mértékegységben legyenek megadva
- Az adatok egyediségére vonatkozó korlátozások ellenőrzése
- A mezőkombinációkban fellelhető inkorrekt értékek felderítése
- Kódolt adatok kódjainak ellenőrzése a rendelkezésre álló kódtáblák alapján
- Azonos tartalmat hordozó, de eltérő adattípusú adatok adattípusainak konverziója
- Az adatok formátumainak ellenőrzése (pl. dátumforma)
- Redundáns adatok megszüntetése
- Alapértelmezett értékekből fakadó problémák feloldása
- Adatkapcsolatok és kardinalitási szabályok ellenőrzése (pl. idegenkulcsok értéke, minimális és maximális kardinalitás)
- Hiányzó adatok kezelése
- Egyéb speciális esetek

Nézzünk néhány példát az előző típusokra az egészségügyi adatok témaköréből! A betegek vizsgálatai kapcsán az egészségügyi információs rendszerek nagy mennyiségű numerikus adatot tárolnak. Az elemzések megkezdése előtt célszerű értelmezési tartományt definiálni minden egyes számszerűen kifejezhető laboradathoz és egyéb numerikus értékhez. Ezáltal például ha a szisztolés vérnyomás felvehető értékeit 50 és 280 Hgmm között határozzuk meg, akkor az ettől kisebb, illetve nagyobb értékek szennyezett adatnak minősülnek. Továbbá, a szisztolés vérnyomás esetében meghatározhatjuk azt is, hogy csak egész értékeket vehet fel, így az összes tizedes számot tartalmazó adatmező szintén nem valós értéknek minősül. Hasonló probléma előtt állunk abban az esetben is, ha például a vércsoport adatokat tekintjük. Mivel a vércsoport összesen 8 különböző értéket vehet fel, ezért érdemes ezeket az eseteket egy listába gyűjteni (vagy törzstáblába), s a felhasználónak csak ezen esetek kiválasztását engedélyezni. Amennyiben a korábbi alkalmazásban ez nem így történt, elképzelhető, hogy a vércsoport mezőben előforduló különböző értékek száma több mint 8. Ekkor tehát újfent meg kell tisztítani az adatainkat. Szerencsés esetben az adatok különbözősége a figyelmetlen adatbevitellel kapcsolatosan csupán a kis és nagy karakterek közti különbségekből fakad (pl.: A+, vagy a+), és nem a karaktertévesztésből. Ugyanis míg az előző esetek könnyen javíthatóak, az utóbbiakban gyakran nem lehet megmondani, hogy mely érték lenne a helyes.

A numerikus adatokat abból a célból is meg kell vizsgálnunk, hogy vajon minden egyes forrásadatbázisban azonos mértékegységben rögzítették-e ugyanazt azt az adatot. Gondoljunk például a hemoglobin szint megadásának módjára. Elképzelhető, hogy az egyik forrásadatbázisban g/dl-ben, míg a másikban g/l-ben adták meg az adatokat. Ekkor megegyezés kérdése, hogy melyik mértékegységet vesszük alapul, s ennek megfelelően az összes más mértékegységben megadott értéket erre a mértékegységre kell átkonvertálni.

Az egyedi értékeket tartalmazó adatmezők, vagy adatmezők kombinációjában található adatok ellenőrzése könnyen megoldható feladat. Ilyen adatellenőrzést kell végeznünk például a betegek TAJ számát tartalmazó mezőre vonatkozóan. Az ismétlődések kiszűrése jelentős mértékben hozzájárulhat a redundáns adatok felderítéséhez is.

Érdekes kérdéskör a tárolt adatok helyességének mezőkombinációs szabályok alapján történő vizsgálata. Vegyük azt az egyszerű példát, hogy a páciens személyes adataiban a neme mező a „férfi” adatot tartalmazza, míg a személyes anamnézisében a jelenlegi állapotára vonatkozóan a „terhes” bejegyzést találjuk. Így biztosak lehetünk abban, hogy a két adat közül valamelyik hibás. Az adatmezők által tartalmazott értékek kombinációs vizsgálatakor érvényességi szabályokat felhasználva vizsgálhatjuk az adatok helyességét.

A kódolt adatok ellenőrzésekor meg kell vizsgálnunk, hogy az adatkódok valós kódok-e, vagyis előfordulnak-e a kapcsolódó törzstáblában, illetve ennek hiányában egyéb rendelkezésünkre álló kódrendszerben. Az egészségügyi adatbázisokban a leggyakrabban előforduló kódok természetesen a betegségek BNO kódjai, illetve a beavatkozások OENO kódjai, de emellett számos megszámlálhatóan véges sok értéket felvehető mező tartalmazhatja kódolva az információt.

Az azonos jelentést hordozó adatok különböző típusú mezőkben történő tárolására vegyük például a kardiológiai ultrahang leleteket. A leletező szoftverek típusától függően elképzelhető, hogy bizonyos adatforrásunkban a mérhető értékeket (pl. bal kamrai diasztolés átmérő, ejekciós frakció) numerikus adatmező rögzíti, míg egy másik adatforrásban a vizsgálati eredmény egyetlen szöveges mezőben tartalmazza ezeket az értékeket. A fennálló ellentmondást célszerű még a migráció előtt feloldani. Amennyiben a szövegmező struktúráltan tartalmazza a mért értékeket, szövegfüggvények segítségével kiszedhetjük belőlük az érdemi adatokat, és adatkonverziót követően numerikus mezőkben helyezhetjük el őket. Azonban ha a szöveges mezőre nem adható meg kis számú séma (maszk), akkor sajnos emberi újraértelmezés nélkül elképzelhető, hogy le kell mondanunk a szövegmezőben tárolt

adatok migrálásáról, s ez által az elemzésekben sem tudjuk majd felhasználni.

Az adatmezők formátumának problémája elsősorban a dátumformákat (pl. születési dátum, elhalálozás dátuma) és a szöveges adatok kis és nagy karakterei közti különbségeket érinti. Ezen problémák feloldása a forrásrendszerekben, illetve a célrendszerben is könnyedén megvalósítható.

A több forrásrendszer adatának összevonásakor előfordulhat, hogy redundáns adatok keletkeznek. Tekintsük például azt az esetet, hogy egy páciens személyi és egyéb adatai több forrásrendszerből is a rendelkezésünkre állnak. Ekkor valamely kód (pl. TAJ szám) alapján azonosítanunk kell ezeket a redundáns adatokat, s meg kell szüntetnünk a többszörös tárolást, hiszen az elemzések során ezek súlyeltolódáshoz vezethetnek. Amennyiben az azonosnak hitt adatok értékükben mégis különböznek, akkor döntenünk kell, hogy mely forrásrendszer adatait tekintjük valósnak, s a másik adatot/adatokat törölni kell a rendszerből.

A forrásadatbázisok adatmezőinek alapértelmezett értékei szintén befolyásolják az elemzéseinket. Ha valamely forrásban egy numerikus adatmező alapértelmezett értéke például „0”, akkor az a matematikai számítások során téves eredményhez vezethet, mivel az összes ki nem töltött érték esetében 0-val fogunk számolni ahelyett, hogy figyelembe sem vennénk az adatot. Bizonyos esetekben a cél vagy forrás adatbázisban könnyen azonosíthatjuk ezeket a szennyezett adatokat, azonban gondban lehetünk például akkor, ha kóros anyagok (pl. antitestek) szintjeinek adataiban 0-t találunk, hiszen nem tudhatjuk, hogy ez az anyag, vagy a vizsgálat elvégzésének hiányát jelenti-e.

A felépített adatbázisunkban természetesen az alapvető adatbázis-kezelési elvárásoknak is eleget kell tennünk, tehát ellenőriznünk kell az elsődleges kulcsok és az idegenkulcsok kapcsolatát (vajon minden vizsgálat ténylegesen hozzárendelhető-e beteghez). Továbbá figyelembe kell vennünk a kapcsolatok kardinalitási viszonyait is (pl. egy betegnél kétszer ugyanolyan szerveltávolító műtétet nem lehet végezni, s ha már eltávolították a szervet, akkor annak nem lehet betegsége).

Az egyes adatmezők hiányos volta jelentős mértékben befolyásolhatja az elemzések eredményességét. Kihagyva az olyan adatokat az elemzésekből, melyek a vizsgált jellemzőket tekintve nem teljesek, előfordulhat, hogy a rendelkezésünkre álló esetszám már nem elegendő a vizsgálat érvényességéhez. További problémát jelenthet az is, hogy a rendelkezésünkre álló adatok téves képet festenek a teljes adatbázisról, mivel nem reprezentatívak az egész populációra nézve. A szakirodalom ezen probléma feloldására számos stratégiát kínál az adatok helyettesítésére vonatkozóan.

A fenti példákból is kiderül, hogy az adattisztítás során vagy lehetőségünk van valóban kijavítani az észlelt hibákat, vagy más megoldás nem lévén kénytelenek vagyunk törölni a rendszerből a hibás adatokat. Alternatív megoldás lehet, hogyha mindenképpen meg szeretnénk tartani az összes adatunkat, de ezek egy része nem értelmezhető számítástechnikai eszközökkel, akkor szakértő emberi munka bevonásával javíttatjuk ki őket. Sajnos az egészségügyi rendszerekben fellelhető és elemzés szempontjából fontos adatok jelentős része strukturáltan szövegformában kerül tárolásra, így ezen probléma megoldása jelenti talán a legnagyobb gondot. Hasonlóan plusz munkaerő bevonást igényel az az eset is, ha elemzéseink további adatokat igényelnek, és ezek az adatok nem állnak elektronikus formában a rendelkezésünkre.

AZ ADATOK ÁTALAKÍTÁSA

A tényleges elemzések megkezdése előtt az adataink a fenti adattisztítási tevékenységeken túlmenően egyéb átalakításokat is igényelnek. Elemzéseinket elvégezhetjük a relációs adatbázisrendszerben tárolt adatokon is, illetve a belőlük épített adattárház adatain is. Azon túl, hogy a második esetben egyéb speciális szerkezeti átalakításokat is végre kell hajtanunk, az elemzések megkezdése előtt célszerű a végfelhasználói igényeknek megfelelően elvégezni a következő tevékenységeket:

- Adatok csoportosítása, rendezése, összegzése, indexelése
- Nézetek, materializált nézetek kialakítása
- Bizonyos folytonos értékeket felvevő adatok átalakítása diszkrét értékekké
- Bizonyos kategorikus adatok numerikus értékekké történő átkódolása
- Új változók felvétele
- A különböző értéket felvevő, de azonos jelentést hordozó adatok megfeleltetése, csoportosítása

Az adatok csoportosításával, rendezésével, indexeléssel, összegzések előzetes számításával, nézetek, materializált nézetek kialakításával az elemző lekérdezések futási idejét tudjuk jelentős mértékben befolyásolni. Egy megfelelően indexelt oszlop a lekérdezések futási idejét sokszor több 10 óráról csökkentheti le akár percekre is. Materializált nézetekben összegyűjtve az adatainkat az elemző algoritmusoknak egyetlen adatobjektumot adhatunk meg input forrásként, s így a futási idő rövidebbé válik, mivel az adatok összekapcsolásával az algoritmusnak már nem kell foglalkoznia.

A folytonos adatok diszkrét csoportokba történő besorolásával az elemzések áttekinthetőségét és az értelmezhetőséget tudjuk javítani. Vegyük például alapul az egyes ionszint értékeket, amelyek a laboreredményekben folytonos értéksorozatot képviselnek. Az elemzések során az elemzést értelmező orvost azonban gyakran nem a konkrét értékek, hanem azoknak a normál értékektől való eltérésének iránya, illetve az eltérés körülbelüli nagysága érdekli. Ez alapján létrehozhatunk különböző kategóriákat. Az elemző algoritmusok futtatása előtt tehát bizonyos folytonos értékeinket diszkretizálnunk kell. Továbbá számos eljárás is megköveteli (például osztályozás problémák esetében a döntési fák alkalmazása), hogy a folytonos értékeket felvevő bemenő adatokat az elemző algoritmus futtatása előtt diszkrét értéktartományokba csoportosítsuk.

Előfordulhat azonban az is, hogy a kategorikus értékeinket kell numerikus adatokkal, vagy értéktartománnyal helyet-

tesítenünk. Ezen átalakításokat szintén az alkalmazandó módszerek korlátozásai teszik szükségszerűvé (például neurális hálók alkalmazása során).

Gyakran előfordul, hogy új, számítható értékek bevezetésével növelhetjük elemzéseink értelmezhetőségét, hatékonyságát. Amennyiben például az adatbázisban tárolt értékek mellett az értékek egymáshoz viszonyított aránya is informatív, akkor célszerű új adatmezőben ezen értékeket is tárolni az elemzések céljából. Így például az LDL koleszterin és a HDL koleszterin mellett a kettő aránya is prediktív értékkel bír.

A különböző értéket felvevő, de azonos jelentést hordozó adatok megfeleltetése kapcsán gondoljunk például arra, hogy hatóanyag szempontjából több gyógyszer azonos hatóanyagúnak minősül, míg az adatbázisban ezek más és más értékeket képviselnek. Éppen ezért célszerű egy új mezőben például hatóanyagaik alapján csoportosítani ezeket az adatokat, s az elemzéseket ezen a magasabb szinten futtatni.

ÖSSZEGZÉS

A fentiek alapján látható, hogy az érvényes eredményeket adó adatelemző tevékenységek egyik elengedhetetlen feltétele az adatok megfelelő szintű előkészítése. Össze kell gyűjtenünk a rendelkezésre álló adatokat, meg kell határozni az elemzésben résztvevő adatok halmazát, ha szükséges ezen adatokat újabb adatokkal kell kiegészítenünk, a meglévő adatokat a téves eredmények elkerülésének érdekében meg kell tisztítanunk, majd az elemző algoritmusok, módszerek hatékony futtatásának érdekében strukturális és egyéb átalakításokat kell rajtuk végeznünk. És mindez még csak az előkészítés folyamata. Ezek után kell majd csak döntenünk, hogy a végfelhasználói igényekhez igazítva az elemzés során mely elemzési módszereket helyezzük előtérbe. Továbbá szakértők értelmezése nélkül elemzéseink eredménye csupán egy újabb adathalmazt eredményez, nem pedig egy esetleges új ismeretet.

Köszönetnyilvánítás A cikkben bemutatott munkát az IKTA 142/2002 számú kutatási projekt támogatta.

IRODALOMJEGYZÉK

[1] Jiawei Han and Micheline Kamber: Data Mining: Concepts and Techniques, The Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann Publishers, August 2000. ISBN 1-55860-489-8

[2] Won Y. Kim, Byoung-Ju Choi, Eui Kyeong Hong, Soo-Kyung Kim, Doheon Lee: A Taxonomy of Dirty Data, Data Mining and Knowledge Discovery 7, 2003

[3] Prather JC, Lobach DF, Goodwin LK, Hales JW, Hage ML, Hammond WE.: Medical Data Mining: Knowledge Discovery in a Clinical Data Warehouse, Proc AMIA Annu Fall Symp. 1997

[4] Tsumoto, S.: Knowledge discovery in clinical databases, Proceedings of the 11th International Symposium on Foundations of Intelligent Systems, 1999.

[5] Stergiani S. Spyrou, Alexander A. Berler, Panagiotis D. Bamidis: Information System Interoperability in a Regional Health Care System Infrastructure: a pilot study using Health Care Information Standards, Proc MIE2003. 2003

[6] M. Last, O. Maimon, A. Kandel: Knowledge Discovery in Mortality Records: An Info-Fuzzy Approach, Medical Data Mining and Knowledge Discovery, Vol. 60, 2001

[7] Papp Á., Márton Á., Adattranszformáció megvalósítási lehetőségek tranzakciós adatbázisok és adattárházak között, NetworkShop 2002 Konferencia, Eger, 2002

A SZERZŐK BEMUTATÁSA

Dr. Fogarassyné Vathy Ágnes A matematika-fizika-számítástechnika szakos tanári diploma megszerzését követően (BDTF, 1995) tanulmányait a Veszprémi Egyetemen folytatta, ahol 1998-ban informatika szakos tanári diplomát szerzett. Jelenleg az ELTE Informatika Doktori Iskolájának hallgatója, abszolutóriumot szerzett 2002-ben. 1998 óta dolgozik a Veszprémi Egyetem Matematikai és Számítástechnikai Tanszékén, 2003 óta egyetemi adjunktus. Jelenleg az IKTA 142/2002 „Intelligens adatelemző központ" projektben szakértő résztvevőként tevékenykedik. Kutatási területei: adatbányászat, adattárházak, az adatbányászati és adattárház módszerek alkalmazása az egészségügyben, adatmodellezés, adatbázisrendszerek.

Dr. Fogarassy György 1995-ben szerzett diplomát a Debreceni Orvostudományi Egyetem Általános Orvosi Karán. 2000-ben belgyógyászatból szakvizsgázott. Az intenzív betegellátás és a nephrológia tárgykörében szerzett gyakorlatot a Szabolcs-Szatmár-Bereg Megyei Önkormányzat Jósa András Kórházában. 2001 óta a Balatonfüredi Állami Szívkórház dolgozója, jelenleg az Intenzív Betegellátó Részlegen tevékenykedik. Érdeklődési területe az arrhythmológia és az invazív kardiológia.