„Big data" az egészségügyi ellátásban

Gergely Tamás, Szöts Miklós
Alkalmazott Logikai Laboratórium, Budapest

A technika fejlődésével az egészségügyben is folyamatosan növekszik a rendelkezésre álló adatok mennyisége. Az egészségügyi problémák összetettebbek lettek, mint valaha, pl. gondoljunk a személyre szabott ellátásra. Ezzel együtt egyre nagyobb az igény a releváns információra, ami gyorsabb hozzáférést tesz szükségessé több és jobb adathoz. Itt jelenik meg a mai információtechnológia egyik legszélesebb alkalmazási kört érintő jelensége, a nagymennyiségű heterogén adattal foglakozó „big data". A cikk bemutatja a „big data" szakkifejezés jelentését, hogy milyen problémákat fed le, és hogy milyen megoldások születnek ezekre a problémákra. Bemutatásra kerül, hogy a kutatási eredmények, alkalmazások milyen lehetőségeket nyitnak az egészségügyi informatikában, valamint hogy a big data jelenség hatására elindult kutatások, fejlesztések hogyan alakíthatják az egészségügyet, hogyan járulhatnak hozzá a páciensek eredményesebb és gazdaságosabb ellátásához.

The evolution of technology in health care creates a growing volume of data with along which an ever-growing demand for information is generated. The health problems became more complex than ever before take, for example, personalized care. This will require a faster access to more and better data. This is it where the phenomenon affecting today's information technology is one of the widest scope of applying large amounts of heterogeneous data dealing with "big data". The article explains the meaning of the term "big data" and refers to the actual problems and to the proposed solutions to these problems. It is shown how the research results and applications open new perspectives in/for health information technology. It is also described how the research and development initiated by the big data phenomenon improves healthcare system and how they can contribute to a more efficient and economical care to the patients.

BEVEZETÉS

Az internet széleskörű elterjedésével véget ért az a korszak, amelyben az adatok strukturált tárolása volt a feldolgozás alapvető feltétele. Óriási mennyiségű adat keletkezik az interneten, például a szociális hálókon, vagy akár a telekommunikációs csatornákon, amelyeknek legalább egy részét célszerű bevonni az adatok közé. Ezzel előtérbe kerül a nemstrukturált adatok kezelése és feldolgozása is. Gondoljunk csak például arra, amikor telefonbeszélgetésekben kell keresni valamilyen terrorcselekmény előkészítésére utaló mintákat. Ebben az új helyzetben nagyon gyorsan keletkezik nagyon nagy mennyiségű adat, nagyon sokféle formában. 2001-ben D. Laney közreadott egy tanulmányt a 3D adatkezelésről [1], ahol a három dimenzió az adatmennyiség, az adatok keletkezésének gyorsasága és az adatstruktúrák különbözősége. Mind a mai napig ezt a 3D modellt használja a szakma a „big data" problémakör leírására. Laney az előbb említett definíciót a következő módon aktualizálta 2012-ben: „a „big data" nagy mennyiségű, nagy sebességű és/vagy nagyon változatos információs eszköz(ök), amely(ek) megköveteli(k) az új feldolgozási formákat annak érdekében, hogy jobb döntéshozatalt, összefüggés kinyerést és folyamat-optimalizálást érhessünk el" [2].

2005-ben a Google jelentős sikert ért el azzal, hogy az influenza terjedését az USA-ban a lekérdezések elemzésével gyorsabban tudta követni, mint a járványügyi hatóság [3]. Azóta a big data divatszó lett. Sok IT eladó és szolgáltató a big data fogalmát kizárólag divatos szakkifejezésként használja az okosabb és kiterjedtebb adatelemzés népszerűsítésére. Ugyanakkor, – ahogy az alábbiakban látni fogjuk, – számos érdekes eredményt köszönhetünk a big data feldolgozásnak. Egyáltalán, mi az, hogy „big data"? Egy jelenség? Problémák gyűjtőneve? Mindenképp létező jelenség, a legkülönbözőbb területeken jelentkeztek ugyanazok a problémák, amelyeket a feldolgozandó adatok rohamos növekedése és különböző strukturáltsága hozott felszínre.

A BIG DATA JELENSÉG

A big data jelenséget a következő ismérvekkel szokták definiálni:

- adatatok nagy mennyisége (a pontos mértékben különbözőek a vélemények és a technika fejlődésével egyre nagyobb adatmennyiséget jelölnek meg),
- az adatbejövetel sebessége meghaladja a szokásos feldolgozás kapacitását,
- több különböző forrásból származnak a feldolgozandó adatok,
- a különböző struktúrájú és a strukturálatlan adatokat (pl. a jól strukturált relációs adatbázisok adatait, a nem strukturált természetes nyelvű szövegeket, jelsorozatokat stb.) kell egyszerre feldolgozni.

Szakértők más jellemzőket is felsorolnak a jelenség meghatározásához, mint pl. az adatok megkérdőjelezhető megbízhatósága.

A fenti meghatározás nagyon bizonytalan. Különböző szakértők más-más adatmennyiségnél látják a big data határ-

át. Az Intel tanulmánya [4] heti 300 terabyte-nyi adatot generáló szervezetekhez köti a big data jelenséget. Azonban a számokat nem érdemes megjegyezni – szinte naponta nő a feldolgozható adatmennyiség. Különben is, a legfontosabb kérdés az, hogy a fent felsorolt ismérvekből hánynak kell egyszerre teljesülnie ahhoz, hogy big data-ról beszéljünk. Például egy kórházi IT rendszerben a betegdokumentációban elsősorban strukturált adatokat találunk. A páciensekről készült szöveges dokumentumok (pl. egy képalkotó vizsgálat eredményének leírása) viszont strukturálatlanok. Big data feladat-e ezek integrált feldolgozása, amikor mennyiségük nem közelíti meg a big data problémákra jellemző méretet, és ugyanabból a forrásból származnak? Hiszen a nem strukturált adatok értelmezése jellemző erre a feladatra, és ez a big data feladatok fontos jellemzője. Ha a betegrekordok feldolgozását összekapcsoljuk egy genetikai adatbázissal, a feldolgozandó adatok mennyisége elérheti a big data határt, de minden adat jól strukturált. Feldolgozásuk során nem jelentkezik a big data feladatokra jellemző adatértelmezési probléma.

A big data fogalmával szembeállítják a small data [25] fogalmát: az adatok jól strukturáltak, egy gépen kezelhetőek, a feladat és megoldási módja átlátható. A small data és a big data jelenség közti különbség bizonytalan, az átmenet köztük folyamatos, ezért megszületett a „medium data" fogalma, amely elválasztja a big data esetét a small data-tól. A medium data [26] kifejezést egyrészt akkor használjuk, ha az adatok feldolgozása már egy gépen nem lehetséges, de mennyiségük még nem kívánja meg, hogy számunkra áttekinthetetlenül, rengeteg gépen elosztva tároljuk. Másrészt akkor, ha az adatmennyiség nem lépné át a hagyományos határokat, de a big data ismérvek közül mások (például az adatok strukturálatlansága) jelentkeznek. Természetesen a határok továbbra is elmosódottak.

A big data jelenséget elsősorban nem az adatbázisok növekedése okozza, hanem az, hogy sok különböző típusú forrásból (pl. web bejegyzés, e-mail, call center hívások felvett anyaga, szenzorok eredménye stb.), sokszor strukturálatlan formában kerülnek az adatok az információs térbe. Így a jelenség okozta problémák többrétűek, amelyek közül a feldolgozás sorrendje szerint a következő problémák a legfontosabbak:

- a gyorsan keletkező, hatalmas adatmennyiség kezelése, begyűjtése, tárolása és elérhetővé tétele, elviselhető időkorlátok alatti kezelése;
- különböző struktúrával rendelkező és nem strukturált adatok együttes feldolgozása,
- hatékony és megbízható információ kivonás az adattömegből.

A big data jelenség azért lett fontos, mert a gyorsan növekvő, strukturálatlan adatmennyiség által generált problémák új technológiákat kívánnak meg, és ilyenek születnek is. A big data jelenség leírására vállalkozó mai kísérletek az új technológiákat a jelenség fontos megkülönböztető jegyének tartják (egy friss és alapos összefoglaló olvasható a definíciókról Ward és munkatársai cikkében [5]).

Fontos, hogy csak az első probléma tartozik kizárólag a big data jelenséghez. Mind a strukturálatlan adatok, mind az információ kivonás problémája jelentkezett már medium data környezetben is; sőt, az egészségügyben felmerülő problémákra épp ezek jellemzőek.

A big data jelenség felgyorsította egy újkeletű (fiatal) tudományág, az adat-tudomány fejlődését. Ennek elsődleges feladata a különböző adattípusok vizsgálatához, valamint az információ¬kinyeréshez megbízható módszerek kiválasztása, illetve kidolgozása.

AZ ADATOK ELÉRÉSE

Az alapvető probléma a nagy adattömeg hatékony kezelése – ezért is kapta a jelenség a big data nevet. Ezzel foglalkoznak legtöbbet, ezen a téren értek el vitathatatlan eredményeket. Kidolgozták a MapReduce párhuzamosítási modellt [6], amelyet a big data problémára létrehozott legtöbb rendszer, így a legelterjedtebb, a Hadoop [27] is alkalmaz. Ugyanakkor van, aki az adattárház technológia felhasználását tartja adekvátnak a probléma kezelésére.

A nagy adattömeg kezelése a leginkább megoldott kérdés, amely a big data jelenséggel kapcsán általában felmerül. Az erre készült rendszerek túljutottak a kísérleti stádiumon, a gyakorlatban használják, sőt, vannak nyílt forráskódú fejlesztések is. Ilyen a már hivatkozott Hadoop, valamint a Cassandra adatbázis kezelő rendszer [28].

Ezzel a problémával nem foglalkozunk részletesebben, egyrészt a fentiek miatt, másrészt mert a hazai IT projektekre nem jellemző a több száz terabyte-nyi adatmennyiség feldolgozása – sőt, a hazai igények nem is követelik meg. Viszont a big data jelenség többi ismérve már jelentkezik medium data szinten is, és az ezek okozta problémák megoldása fontos a magyar egészségügynek is.

AZ ADATOK ÉRTELMEZÉSE

Jelentős probléma, hogy a nem strukturált adattömegből (szenzorok jelsorozatai, audió és video fájlok, természetes nyelvű szövegek) kiválogassuk a feladatnak megfelelő adatokat. Természetesen a „nem strukturált" jelző IT szempontból értendő: a természetes nyelvű szöveget például az ember számára jól strukturálja a nyelv grammatikája.

Míg a strukturáltan tárolt adatok esetén a tárolás struktúrája szerint lehet az adatokhoz hozzáférni, a strukturálatlanul előforduló adattömeg esetében magából az adatból kell eldönteni, hogy releváns-e a feladathoz, és ha igen, a feladat szempontjából milyen értékes információt hordoznak. Néha elég egy adott elem előfordulását megtalálni az adatban, például a vélemény analízis (sentiment analysis) esetében bizonyos szavak előfordulásait keresik természetes nyelvű szövegekben. Azonban általában a helyzet sokkal bonyolultabb, kiterjedt kutatás szükséges az adatfolyam értelmezéséhez. Ilyen kutatások nem kötődnek szorosan a big data jelenséghez, már sokkal előbb elindultak, és jelentős eredmények születtek. Itthon is folyik ilyen kutatás, például a természetes nyel-

vű szövegekből való információ-kivonás területén elért eredményekről számol be [7], [8]; a beszédfeldolgozás területéről pedig érdemes megtekinteni a http://alldio.eu honlapot.

Bár a problémák jelentősen különböznek a vállalt feladattól és az adatfolyam jellegétől függően, a feladat általánosan megfogalmazható: az értelmezés azt jelenti, hogy az adatokat fogalmakhoz kell kapcsolni. Más szavakkal: szemantikus elemzés végzendő. Kifejezetten egy feladatra irányuló fejlesztésnél esetleg nem jelentkeznek a szemantikus feldolgozás módszerei, eszközei. Azonban, ha a strukturálatlan adatok egy fajtájának értelmezését általánosabb szempontból akarjuk megvalósítani, ezek alkalmazandóak; első sorban a különböző fogalmi hálók. A World Wide Web Consortium (W3C) alapos felmérést készített arról, hogy a szemantikus feldolgozás milyen mértékben használatos. Ez a felmérés és [9] összefoglalja a szemantikus feldolgozással kapcsolatos jelenlegi kihívásokat és ezekre adható néhány fontosabb megoldást, valamint összefoglalja a jövő kihívásait. A szemantikus feldolgozás egyik fontos iránya az ontológiatechnológián alapul. Ez az irányzat az ontológiatechnológia összes válfaját alkalmazza, különös hangsúllyal az automatikus, gépi tanuláson alapuló ontológiaépítést is. A big data jelenség külön kihívást is állít az ontológia technológia elé: általában nem azonos megbízhatóságú adattömegekkel kell dolgozni. Ez az ontológiatechnológiának egy új irányát hozta létre, a bizonytalan megbízhatóságú adatokból való ontológiagenerálást.

Ne feledkezzünk el arról, hogy sokszor célszerű egyszerre különböző struktúrájú, ill. strukturálatlan adatok közt keresni. Az erre szolgáló megoldásokat is ki lehet fejleszteni és alkalmazni lehet medium data környezetben, ilyen például a „Remote Accessibility to Diabetes Management and Therapy in Operational healthcare Networks" megnevezésű EU projektben (a projekt a diabetesz kezelésére dolgozott ki információtechnikai eszköztárt) született szemantikus alapú megoldás, amely egyszerre keres egy páciens elektronikus rekordjában és kapcsolódó természetes nyelvű dokumentumokban.

ANALITIKAI MEGOLDÁSOK

A rendelkezésre álló hatalmas adattömeg és ennek szinte hihetetlen növekedése csak akkor jelent kihívást részünkre, ha olyan értékelhető információt vonhatunk ki belőle, amelyre szükségünk is van. Ehhez nemcsak az adatforrások hatékony elérése szükséges, hanem azok hatékony, megbízható feldolgozása is; sőt, ez az elsődleges szempont. Tehát az információ-kivonás módszerei – amelyek általában különböző analitikai módszerek – döntik el a big data adatfeldolgozás eredményességét. Az analitika az adatok szisztematikus feldolgozását jelenti releváns elvárások tükrében valamilyen analitikai diszciplína segítségével (pl. statisztika valamilyen modellje alapján). A feldolgozás eredménye vagy a felhasználónak, vagy magának a rendszernek segít egy tényalapú döntés meghozásában.

A big data feladatoknál alkalmazott analitikai megoldások skálája igen széles: az egyszerű statisztikától bonyolult struktúrákig (pl. Bayes hálók) minden használatos. Bármilyen problémát is okoz ezeknek az analitikai megoldásoknak a big data feladatokon való alkalmazása, a módszerek azonosak a medium data esetén használtakkal. Azonban az alkalmazás feltételei, módjai big data feladat esetén változnak. A big data jelenséggel bekövetkezett forradalmi változás az, hogy az adatok elemzése közvetlen gyakorlati eredményeket szolgáltat az elméletet kikerülve; sőt a rendszerek az adatok elemzése alapján tanulják működésüket. (A tanuló programok alkalmazása természetesen nem teljesen új jelenség. Számos területen a big data jelenségtől függetlenül is tanul a számítógépes rendszer, például a természetesnyelv-feldolgozás területén.) Pontosabban: az adatokban felfedett összefüggéseket elfogadják, anélkül, hogy oksági viszonyok megállapítására törekednének. A gyakorlat elmélet feletti győzelmét látják ebben. Kérdés azonban, mennyire megbízható az analitikus módszerek eredményeinek alkalmazása anélkül, hogy az aktuális területről való elméleti tudásunk segítségével értékelnénk. A cikkünk elején említett, az influenza terjedését követő Google Flue Trends rendszer másodszori alkalmazása nem volt sikeres, – és még azt sem lehet tudni, hogy miért nem.

Az analitikus módszerek alkalmazása területén is vannak problémák, mégpedig jelentkeznek ugyanazok is, amelyeket már a klasszikus analitikus megoldásoknál is megismertünk. Ezek a big data jelenség esetén nem tűnnek el, csak súlyosabbak lesznek, kevésbé ellenőrizhető az analízis folyamata. Az alapvető kérdés a mintavétel és annak kritériumai. Viktor Mayer-Schönberger, az Oxford's Internet Institute professzora, a Big Data c. könyv egyik társszerzője szerint a valóban nagy adathalmaz definíciója "N = minden", ahol N a mintahalmaz számossága [10]. Azonban a rendelkezésre álló adatok sosem fedik le a teljes valóságot. Például az internethasználat különböző tevékenységei, amelyek nyomai big data jelenség tipikus adatforrásai, nem tükrözik a teljes lakosságot, mivel különböző végzettségű, korú és lakóhelyű emberek eltérő mértékben különböző módon használják az internetet. A hagyományostól való eltérés az is, hogy a big data feldolgozás esetén nem rekonstruálható a mintahalmaz, amely alapján az eredmény megszületett. Ugyanis az alkalmazás újrafuttatása még ugyanazokon az adatforrásokon sem fogja pontosan ugyanazokat az adatokat használni. Ugyanazokon az adatokon más alkalmazást, más analitikus módszert futtatni lehetetlen. Mindez persze nem azt jelenti, hogy a big data jelenség adathalmazain nem szabad analitikus módszereket alkalmazni, hanem csak óvatosságra int. Vannak esetek (például marketing célból végzett feldolgozás a szociális hálón), amikor a nem megfelelően ellenőrzött összefüggés is felhasználható. Azonban sok területen, például az egészségügyben, óvatosan kell eljárni. A nagy adathalmazon kimutatott összefüggéseket hipotézisekként kell kezelni, egyéb módszerekkel ellenőrizni.

A big data jelenség azonban ténylegesen lendületet adott az analitikai eszközök használatának: megszületett a prediktív analitika, amely lehetővé teszi az előrejelzést [11]. A prediktív analízis esetében nemcsak bizonyos jellemzők együttesét keresik az adathalmazban (ezeket nevezik prediktoroknak), hanem ezek alapján megbecsülik, hogy milyen valószínűséggel fog valamilyen esemény bekövetkezni [30].

Az analitikus módszerek használata big data problémák esetén számos átütő sikert hozott. Egyik nagy visszhangot kiváltó siker az IBM által kifejlesztett Watson rendszerhez kötődik. A rendszer első lett a Jeopardy! kvíz versenyen úgy, hogy két nagyon jó kvíz játékost vert meg [31]. A sikerek bizonyítják, hogy komolyan kell venni mind a big data jelenség problémáit, mind az ezekre kidolgozott szoftver megoldásokat. Azonban a sikerek alapján sokan kétségbe vonják a tudományos módszerek használatának szükségességét, az analitikus módszerek direkt alkalmazásával ezeket helyettesíthetőnek gondolják. Kritikusan kell kezelni ezeket a sikerektől elvakult véleményeket, amilyent például egy big data-ról szóló könyvben olvashatunk, ahol a szerzők szerint „a kauzalitást nem vetjük el, de eltávolítjuk arról a piedesztálról, amelyen a jelentés első számú forrásaként foglal helyet" [12].

BIG DATA MINT EGÉSZSÉGÜGYI ERŐFORRÁS

Az adatokat már a kilencvenes években is az egészségügy fontos erőforrásának tekintették. Azóta az egészségügy területére is jellemző az egyre bővülő adattömeg. Ez egyre gyorsuló mértékben növekszik – és így itt is szembetalálhatjuk magunkat a big data jelenséggel. Ugyanakkor ennek az adattömegnek hatékony és megbízható feldolgozása szükséges a magasabb minőségű és alacsonyabb költségű ellátás eléréséhez. A big data problémák itt is jelentkeznek, az ezekre született megoldások az egészségügyben is jelentős kutatási és gyakorlati eredményt hozott. A következőkben utalunk néhányra, további érdekes eredményekről Hersh könyvében olvashatunk [13].

Ma már az ellátás során rengeteg adat kerül rögzítésre és megőrzésre adatbázisokban. Ha az egyes páciensek rekordjaiból esettárakat alkotunk, számos hasznos összefüggést tárhatunk fel. Például a 2-es típusú diabétesz gyógyszerelésének hatását vizsgálták az Optum Labs adatbázisán, több mint 37 000 páciens adatain [32]. Az eredmények szerint több szempontból (vércukorkontroll, életminőség, élettartam) nem volt különbség, de a sulfonylurea használata esetén a költségek alacsonyabbak voltak, és később kellett a pácienseket átállítani inzulinra.

Az Egyesült Államokban a HMO keretében létrehoztak egy virtuális adattárházat (VDW). Az ehhez kapcsolódó kutatóhelyek szabványos módon igényelhetnek adatokat kutatásaikhoz. A VDW különböző EMR adatbázisokból beszerzi az anonimizált adatokat, ezeket egységes alakra hozva bocsájtja a kutatóhelyek rendelkezésére [14]. Egy érdekes eredmény: egy, a VDW-re támaszkodó kutatás megállapította, hogy összefüggés van a gyermekkori elhízás és a terhesség alatti hyperglycaemia közt [15].

A páciensre vonatkozó adatok jelentős részét természetes nyelvű dokumentumok tárolják [16], ezért fontos kidolgozni az adatbázisrekordok és a természetes nyelvű dokumentumok integrált kezelését. Az adatok nagy részét azonban nem is őrizzük meg. Ilyenek például a szenzorok által szolgáltatott adatfolyamok. Ezek felhasználása ugyancsak értékes összefüggéseket tárna fel. Például a Torontói Gyermekkórházban vizsgálták az újszülöttek monitorozásából kapott adatokat (másodpercenként több mint 1000 különböző fiziológiai mérési eredményt). Ezek alapján a kutatók algoritmust állítottak össze annak jelzésére, hogy egy csecsemőt veszélyeztet-e fertőzés [17].

Sajnos sokszor adat sem lesz a páciensek állapotjellemzőiből és azok körülményeiből. Egy krónikus betegségben szenvedő páciens – jó esetben – havi, kéthavi rendszerességgel felkeresi orvosát, s akkor történik adatfelvétel. Azonban közben számos esetben történik vele valami, amely befolyásolhatja a kezelést, s gyakran akár azonnali útmutatást kívánna. A jelenlegi technológiák felhasználásával ez ma már nem lehetetlen. Példa: a Ginger.io mobil alkalmazást fejlesztett ki, amely segíti a páciensek (pl. cukorbetegek) helyes életmód választását [18]. Az alkalmazás a páciens pillanatnyi tevékenységének, állapotának számos adatát közvetíti a központba, ahol a rendszer az adatokat releváns tudásbázisokkal összevetve tanácsokkal látja el a pácienst. Természetesen ennél egyszerűbb eszközök is sokat segítenének.

Szót kell ejteni a szociális hálókról, amelyek a big data paradigma tipikus adatforrásainak számítanak. Ma is számos az egészségi állapotról szóló bejegyzés található rajtuk. Ahhoz, hogy a big data paradigma eszközeit alkalmazzuk, két akadályt kell leküzdeni. Az első az, hogy fokozottan jelentkezik a személyiségi jogok problémája, amelyet sokan általában a big data feldolgozás problémájának tartanak. A legtöbb szóba jöhető adat egy páciens adata, amelyhez való hozzáférést szigorú etikai és törvényi szabályok korlátozzák. A big data jelenség lehetőségeinek kihasználása érdekében fontos lenne ezek átvizsgálása, az adatokhoz való jogok további szigorú biztosítása mellett. Olyan EMR rendszereket lehetne kidolgozni, amelyek lehetővé tennék az anonimizált hozzáférést. A második akadály a hozzáférés. Jelenleg még azt sem lehet biztosan tudni, hogy egy személy egészségügyi adatai hol érhetők el. E két probléma szorosan összefügg. A problémák megoldására világszerte több jelentős állami (az USA-ban magán) projekt van folyamatban, és több helyen értek el eredményeket. Például mindkét problémát megoldja az HMO virtuális adattárháza [14]. Hazánkban is elindult egy projekt az egészségügyi kooperatív tér [19] megvalósítására, amely az adatok fellelhetőségét segítené.

AZ INFORMÁCIÓFELDOLGOZÁS

Eddig a kérdéskört az adatok szempontjából tekintettük át. Néhány szót kell ejtenünk a paradigma legfontosabb vetületéről is, az információfeldolgozás módszereiről – azaz az alkalmazható analitikai módszerekről. Az alkalmazott módszerek kérdésében nincs lényeges különbség az alkalmazás területei közt. Az egészségügyi alkalmazások analitikai módszerei a legegyszerűbbektől a legbonyolultabbakig terjednek; például egy, az EMR-be ágyazott Bayes-féle hálózaton alapuló modell alkalmazásával egy kórházban előre jelezhető volt a felfekvéses esetek kialakulása, így ezeket tizedére sikerült csökkenteni [20]. Az egészségügyi alkalmazások közt jelentős szerepet játszik a prediktív analitika, a hivatkozott kutatás ennek is jel-

lemző példája. Azonban a big data paradigma kritikáját az egészségügy területén nagyon komolyan kell venni. Az analitikáról szólva már felhívtuk a figyelmet a statisztikai módszerek meggondolatlan használatának problémáira. Amikor az egészségről van szó, óvatosnak kell lennünk. A feltárt összefüggéseket, mint hipotéziseket lehet használni, s azokat a szokott szigorú módszerekkel igazolni.

ÉRTÉKTEREMTÉS AZ EGÉSZSÉGÜGY SZÁMÁRA

Végül a legfontosabb kérdés: mit nyerhet az egészségügy? Az eddigi példák is mutatják, hogy jelentős segítséget jelentenek a big data paradigmához tartozó megoldások, de érdemes áttekinteni mindezt az egészségügy szintjein.

Országos szinten (az egészségpolitika és a népegészségügy területén) jelenleg az ellátó helyek által szolgáltatott jelentéseket dolgozzák fel. Big data feldolgozásról akkor beszélhetnénk, ha az egyes adatokat integrálás nélkül, közvetlenül érnék el az IT rendszerek, – ehhez a már hivatkozott kooperatív tér szükséges. Az országos szinthez hasonló feladatok merülhetnek fel intézményi szinten is. Mindkét szinten a feladat az ellátás minőségének emelése mellett az ellátás gazdaságosságának biztosítása. Több helyen értek el eredményt például a kórházi újrafelvételek csökkentésében – [21] számol be ilyen eredményről.

Az igazán izgalmas terület a klinikum, amikor egy páciens kezelése során történik IT rendszerek felhasználása. A big data paradigmán belüli kutatások nagyon sok olyan eredményt hoznak, amelyek felhasználhatóak az ellátás során. Az eddig hivatkozottak is ilyenek. A legfontosabb kutatási irányok, amelyek megváltoztathatják az ellátás módját:

- Terápiák kimenetelének kutatása, azaz annak eldöntése, hogy bizonyos kórkép esetén melyik terápia a legkedvezőbb. Ilyen például az Optum Labs adatbázisán elvégzett fent idézett kutatás.
- Prediagnózisra irányuló kutatások, amelyek azt kívánják felfedni, hogy bizonyos betegségek milyen jelekből állapíthatók meg még az előtt, hogy a tradicionális eljárás diagnózist adna. Ilyen az újszülöttek fertőzésveszélyét előre jelző rendszer [17], vagy az onkológiai esetek prediagnózisa [13].
- Krónikus betegek telemedicinális megfigyelése, amelyből lemérhető mennyire követik a terápiás előírásokat, mikor szorulnak életmódi tanácsadásra, mikor kell orvoshoz fordulniuk.
- A páciensre szabott kezelés, például a génadatbázisok felhasználásával.

A big data jelenség eszközei nemcsak a kutatásban lennének használhatóak, hanem közvetlenül az ellátás feladataiban is. Ilyen feladat például problémás páciensehez való hasonló eset keresése. Előre preparált esettárakban keresni medium data feladat lehet (sőt, egy esettár esetén small data). Azonban ha az EMR-ekben és a releváns orvosi dokumentumokban (esetleg kiegészítve a web bejegyzésekkel) keresünk, tipikus big data feladattal állunk szemben. Ehhez nemcsak a természetes nyelvű ellátási dokumentumokba való keresést kell megoldani, hanem az adatbázissal való integrált keresést is. Célszerű a keresésbe bekapcsolni az irodalomban (protokollok, cikkek stb.) való keresést is. Ezt teszi az IBM Watson rendszere is, amikor az onkológiai esetek ellátását segíti [33].

Az egészségügyi adatok hasonlóságai alapján patofiziológiai minták („disease signatures") kialakítását tűzi ki célul [22]. Ezek megkönnyítenék a felgyülemlő páciensadatok felhasználását, így hatékony patofiziológiai diagnózis megállapítását tennék lehetővé. Hasonlóképpen lehetne felhasználni a szociális médiát arra, hogy a hasonló problémával küzdő emberek megtalálhassák egymást, és tapasztalataikat megoszthassák.

A fenti összefoglalásban nem akartuk elválasztani a medium és a big data feladatokat, eredményeket. Már csak azért sem, mert nehéz pontosan meghúzni a határvonalat, illetve, mert a medium data feladatokra kidolgozott megoldások használhatóak lehetnek a big data paradigmán belül is.

ÚJ PARADIGMA: A KOGNITIV SZÁMÍTÁSTECHNIKA

Ginni Rometty, az IBM vezérigazgatója a Watson rendszer sikereire alapozva a számítástechnika új korszakát vetíti előre [34] . Nem azért, mert a Watson rendszer hatalmas dokumentumtömeget képes átvizsgálni keresés közben. Hanem egyrészt azért, mert a rendszer teljesítménye a szövegek értelmezésében megközelíti, illetve bizonyos szempontokból el is éri az ember teljesítményét; másrészt pedig a rendszer működése során tanulja, hogy hogyan működjön pontosabban, hatékonyabban. A rendszer kifejlesztéséhez nem csak a számítógépes nyelvészet eszköztára volt szükséges, hanem olyan technológiák is, amelyek képessé teszik a rendszert következtetések levonására és állandó tanulásra. Már az értelmezéshez is szükséges volt egy egyszerű időlogika, mivel ez szükséges az állításokból kiszűrt tények időbeli elrendezéséhez.

S ezzel elérkeztünk egy újabb bonyolultsági fokhoz: a kognitív informatikához, ahol a rendszer már megvalósítja a régi álmot: eszközből problémamegoldó társ lesz. A kognitív informatika kialakításához három fontos területen kell erős technológiával rendelkezni:

- A rendszerrel való kapcsolatban túl kell lépni a szigorúan formalizált nyelveken. Elsőrendű követelmény, hogy a természetes nyelvet értse a rendszer. Azaz hogy megközelítse azt a szintet, ahogy mi értjük. A szövegekből olyan belső reprezentációt állítson elő, amely a téma szempontjából fontos információt reprezentálja – legalább az aktuális tématerületen. Azonban tovább is léphetünk: nemcsak a rendszer és a felhasználó közti kapcsolat lehet fontos, hanem a rendszer és a környezete közti is. Az input jöhet több csatornán párhuzamosan, ez tipikus big data jelenség. Ekkor nemcsak azt várjuk el, hogy az egyes adatfolyamok elemeit értelmezni tudja a rendszer, hanem azt is, hogy az egymásra utaló jeleket összekapcsolja, és az input által leírt szituáció reprezentációját előállítsa. Tehát a különböző inputcsatornákon (audio, video, text stb.) beérkező adatfolyamot összehangolt módon, integráltan, egységes szemantikus elmélet alapján kell feldolgozni. Ezt a

jelenséget nevezik szemantikus alapú adatfeldolgozásnak.

- A rendszernek képesnek kell lennie bonyolult következtetési feladatok megoldására. Több különböző következtetési rendszert kell kezelnie, és rendelkeznie kell feladatmegoldó stratégiával. Ilyen rendszer kiépítéséhez jól megalapozott logikai keret kell [23].
- A rendszernek tanulnia kell, mégpedig úgy, hogy egyrészt tudását állandóan bővíti, másrészt az elvégzett feladatmegoldások tapasztalatait hasznosítja, – azaz tanulnia kell feladatmegoldó stratégiát is. Ez megint csak a big data technológia egyik kihívása, a Watson rendszerrel kapcsolatban is ezt emelik ki.

A fenti technológiák kidolgozása, integrálása még ma is jelentős kutató-fejlesztő munkát igényel, de a big data jelenség problémáira kidolgozott megoldások már a kognitív informatika irányába mutatnak.

ÖSSZEFOGLALÁS

Az egészségügyben a big data megoldásai különösen két terület számára jelentenek fontos mozgató erőt. Az egyik a klinikai és transzlációs informatika, amelynek feladata a kutatási adatok információvá, illetve tudássá történő alakítása és ennek a tudásnak a betegek ellátásában történő felhasználása. A klinikai és transzlációs Informatika magában foglalja az orvosbiológiai informatikát, a klinikai kutatás informatikájának egyes részterületeit, a transzlációs bioinformatikát, képalkotó informatikát, valamint ezek kapcsolódásait (közös részeit) a klinikai informatikával és a népegészségügyi informatikával.

A másik fontos terület az ellátás perszonalizálása, amelynek keretében a cél a személyre szabott, prediktív, megelőző, és a részvételre építő egészségügyi ellátás megteremtése. A perszonalizált ellátás az ellátottak jóllétére és a megelőzésre fókuszál. A perszonalizált ellátás központi eleme a személyre szabott, perszonalizált ellátási, illetve egészségmegőrzési terv, amelynek fontos eleme a kockázat menedzsment terv. Ez utóbbi feltételez különböző monitorozási adatok együttes kezelését és ezekből megfelelő információ kinyerését. Megjegyezzük, hogy a perszonalizált ellátás betegközpontú klinikai döntéstámogatást igényel, amelyhez új adatkezelési stratégiákat kell létrehozni a hatalmas mennyiségű, potenciálisan érzékeny egészségügyi adatok kezelésére.

E célok eléréséhez azonban nem elsősorban a nagy adatmennyiség kezelése a kulcs, hanem a strukturálatlan adatok értelmezése. Ezért medium data környezetben kidolgozathatóak a technológiák; sőt, számos esetben medium data környezetet igényelnek a konkrét feladatok.

Ha távolabb tekintünk, az egészségügy új struktúrája körvonalazódik. Hagyományosan a tevékenységi térben az ellátás központosított intézményekben, elsősorban kórházakban történik. Az információ azonban a szükségletekhez mérve széttagolt, a különböző helyeken születő információt rögzítő adatok más helyeken nem, vagy csak nehezen férhetőek hozzá. Ez plusz tevékenységeket jelent az ellátásnak, ezért a szükségesnél költségesebb, ronthatja az ellátás minőségét, a páciensnek meg kényelmetlen. Az információtechnológia fejlődése feszegeti ezeket a határokat, és a big data problémákra kidolgozott megoldások tisztábban kirajzolnak egy jövőt, ahol a tevékenységek szétszórtan történhetnek (az otthoni ellátás dominánssá válása), és az információs tér központosított lesz, azaz a keletkező adatok ott lesznek elérhetőek, ahol szükségesek [24].

IRODALOMJEGYZÉK

[1] Laney, D.: 3D Data Management: Controlling Data Volume, Velocity, and Variety, 2001 http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

[2] Laney, D.: The Importance of 'Big Data': A Definition, Gartner, Retrieved, 21 June 2012.

[3] Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., Brilliant, L.: Detecting influenza epidemics using search engine query data, Nature, 457, 1012-1014, 2009

[4] Intel Peer Research on Big Data Analysis. http://www.intel.com/content/www/us/en/big-data/data-insights-peer-research-report.html.

[5] Ward, J.S., Barker, A.: Undefined By Data: A Survey of Big Data Definitions

[6] Dean, J., Ghemawat, J. [2004]: MapReduce: Simplified Data Processing on Large Clusters http://static.googleusercontent.com/media/research.google.com/es/us/archive/mapreduce-osdi04.pdf

[7] Szőts M., Csirik J., Gergely T., Karvalics L.: MASZEKER: projekt szemantikus kereső rendszer kidolgozására, MSzNy 2010, http://www.inf.u-szeged.hu/mszny2010

[8] Szőts M., Gyarmathy Zs., Simonyi A.: Frame-szemantikára alapozott információ-visszakereső rendszter, MSzNy 2013, http://www.inf.u-szeged.hu/mszny2013

[9] Jeong, R. J., Ghani, I.: Semantic computing for Big Data Approaches, Tools, and Emerging Directions (2011-2014), KSII Transactions On Internet And Information Systems, Vol. 8, No. 6, June 2014 2022-2042

[10] Harford, T.: Big data: are we making a big mistake? Financial Times http://www.ft.com/intl/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html#axzz3GE25hO6A

[11] Matlis J.: QuickStudy: Predictive Analytics, 2006 http://www.computerworld.com/article/2554079/business-intelligence/predictive-analytics.html

[12] Mayer-Schönberger, V., Cukier K.: Big Data, A Revolution That Will Transform How We Live, Work, and Think John Murrays (Publishers), 2013

[13] Hersh, W. R.: Healthcare Data Analytics. in: Hoyt RE, Yoshihashi, A, Eds. Health Informatics: Practical Guide for Healthcare and Information Technology Professionals, Sixth Edition, Pensacola, FL, Lulu.com, 2014
[14] Ross, T.R., Ng, D., Brown, J.S., Pardee. R.: The HMO Research Network Virtual Data Warehouse: A Public Data Model to Support Collaboration, eGEMs, EDM Forum, Mar 2014, Vol. 2, Iss. 1 http://repository.academyhealth.org/egems/vol2/iss1/2/
[15] Hillier TA, Pedula KL, Schmidt MM, Mullen JA, Charles MA, and Pettitt DJ. Childhood obesity and metabolic imprinting: the ongoing effects of maternal hyperglycemia, Diabetes Care, 30: 2287-2292., 2007
[16] Jensen, P.B., Jensen, L.J., Brunak, S.: Mining electronic health records: towards better research applications and clinical care, Nature Reviews Genetics http://www.nature.com/nrg/journal/v13/n6/abs/nrg3208.html, 2012
[17] Infoway, Big Data Analytics in Health. Canada Inforoute, Canada Health Infoway Inc. https://www.infoway-inforoute.ca/index.php/resources/technical-documents/emerging-technology, 2013
[18] Groves, P., Kayyali, B., Knott, D., Kulken. S. van: The 'big data' revolution in healthcare, Center for US Health System Reform Bussiness Technology Office, Jan, 2013 http://www.mckinsey.com
[19] Szathmáry B.: A kórházak szerepe a Kooperativ Térben, IME – Az egészségügyi vezetők szaklapja, XII. évf. 1. sz. pp 57-58, 2013
[20] Cho I, Park I, Kim E, Lee E, Bates DW.: Using EHR data to predict hospital-acquired pressure ulcers: A prospective study of a Bayesian Network model, International Journal of Medical Informatics, 82: 1059-1067., 2013
[21] Gilbert P, Rutland MD, Brockopp D.: Redesigning the work of case management: testing a predictive model for readmission, American Journal of Managed Care, 2013. 19 (11 Spec No. 10): eS19-eSP25., 2013 http://www.ajmc.com/publications/issue/2013/2013-11-vol19-sp/redesigning-the-work-of-case-management-testing-a-predictive-model-for-readmission.
[22] Morley-Fletcher E.: Big Data Healthcare – An overview of the challenges in data intensive healthcare, Networking Session on Big Data in Healthcare at ICT 2013. http://www.lynkeus.eu/big-data-healthcare/
[23] Anshakov, O, Gergely T: Cognitive Reasoning – a Formal Approach, Springer 2010
[24] Gergely T.: Integrated care space for chronic diseases, E-Health Week, Budapest, HIMSS, 2011
[25] http://whatis.techtarget.com/definition/small-data
[26] http://highscalability.com/blog/2014/12/17/the-big-problem-is-medium-data.html
[27] http://hortonworks.com/blog/apache-hadoop-yarn-concepts-and-applications/
[28] http://cassandra.apache.org/
[29] http://www.w3.org/2001/sw/sweo/public/UseCases/
[30] https://www.priv.gc.ca/information/research-recherche/2012/pa_201208_e.asp#_ftn6
[31] http://researcher.watson.ibm.com/researcher/view_group_subpage.php?id=2160
[32] http://content.healthaffairs.org/content/33/7/1187.short
[33] http://www.informationweek.com/software/information-management/ibms-watson-could-be-healthcare-game-changer/d/d-id/1108608?
[34] http://www.informationweek.com/software/information-management/ibm-ceo-rometty-shares-vision-of-big-data-era/d/d-id/1108992?

A SZERZŐK BEMUTATÁSA

Gergely Tamás a műszaki tudomány kandidátusa, a matematikai tudomány doktora, címzetes egyetemi tanár, az Orosz Természettudományi Akadémia tagja, az Alkalmazott Logikai Laboratórium ügyvezető elnöke. Kutatási területei: matematikai logika, számítástudomány, mesterséges intelligencia, nagy komplexitású rendszerek modellezése, intelligens adat-vizsgálati módszerek. Alkalmazási területek: intelligens kooperatív rendszerek, döntéstámogató rendszerek, tudásmenedzsment rendszerek és alkalmazásuk az egészségügyi informatikában, valamint rendszer orvosbiológia. Több mint 160 közleménye és 8 könyve jelent meg.

Szőts Miklós mérnök, matematikus, matematikai tudomány kandidátusa, 1986 óta az Alkalmazott Logikai Laboratórium kutatója. Számos hazai, nemzetközi és ezen belül európai uniós (FP5, FP6, FP7) egészségügyi IT projektben vett részt. Szakmai területe a természetes nyelv feldolgozása, a virtuális egészségügyi ellátás, valamint az EHR rendszerek (openEHR, kétszintű adatmodell).