A BigData lehetőségei a közlekedésben

A közlekedési informatikai rendszerek, hasonlóan más korszerű informatikai rendszerekhez óriási mennyiségű adatot generálnak. Ezek felhasználása a tervezési, irányítási folyamatokban ma csak részben történik meg. A cikk bemutatja a BigData eszközrendszerét, a felhasználás folyamatát, és valós példákon keresztül pedig azt, hogy az adatok célszerű csoportosításával milyen új ismeretekhez lehet hozzájutni.

DOI 10.24228/KTSZ.2017.4.4

Dr. Horváth Richárd
Széchenyi István Egyetem Logisztikai és Szállítmányozási Tanszék
e-mail: horvath.richard@sze.hu

1.BEVEZETÉS

A nagy méretű adatbázisok, adathalmazok puszta adattemetők, – amennyiben az adatokat tárolják, megőrzik– mindaddig, amíg fel nem dolgozzuk azokat. A hatalmas adathalmazok feldolgozása, a hasznos összefüggések kinyerése nagy kihívást jelent [1]. John Naisbitt mondta: „Miközben belefulladunk az információba, tudásra éhezünk". Az utóbbi évek trendje, hogy egyre nagyobb mennyiségű információt állítunk elő. 2015-ös adatok szerint a 2013-2015 közötti időszakban annyi adatot állított elő a világ, mint azt megelőzően összesen. A technika fejlődésével egyre olcsóbbá váló adatrögzítő eszközök, érzékelők, szinte kínálják a lehetőséget arra, hogy gyűjtsünk minél több adatot. Felmerül a kérdés azonban, hogy miért, ha ezzel a későbbiekben nem történik semmi. Sok esetben az adatgyűjtés nem a későbbi felhasználás érdekében történik, hanem az operatív beavatkozások, döntések támogatása, ill. megalapozása miatt. A mérnökök, adatfelhasználók felelőssége, hogy meghúzzák a határt a szükséges és a felesleges adatok között, hogy csak az értékes adatokat tárolják.

2. BIGDATA – NAGY MENNYISÉGŰ ADAT A KÖZLEKEDÉSBEN

A BigData kifejezés értelmezése könnyebbé válik, ha az irodalmakban [2] használatos 4V meghatározást használjuk:

- nagy mennyiség (Volume),
- időben gyorsan változó (Velocity),
- változó formátumok (Variety),
- adat minőségének váltakozása időben (Veracity / Variability).

A nagy mennyiség megfogalmazás sok esetben az ember számára már nem értelmezhető mennyiséget jelent, hiszen egy mozifilmet tartalmazó 4 Gbyte-os DVD még értelmezhető, azonban egy korszerű Ford Fusion gépjármű óránként generál 25 Gbyte adatot, míg egy transzatlanti útvonalat teljesítő Boing-777-es 30 Tbyte-nyi adatot generál. Mindkét esetben látható, hogy ekkora mennyiségű adat tárolása egyrészt lehetetlen bármilyen olcsóvá is válik a tárolás, másrészt felesleges is, hiszen az adatok jó része az adott pillanatban érvényes volt, utána pedig már nincs jelentősége.

A BigData-val kapcsolatos alkalmazások kiinduló pontja az adatforrás. Két nagy csoportot lehet megkülönböztetni az analóg és a digitális adatokat. Analóg adatról beszélünk a videófelvételek és a hangfelvételek esetén. A közlekedésre inkább a digitális adatok a jellemzők, mint helymeghatározó rendszerből származó adatok, idő adatok, eszközök által generált adatok, érintéses és érintés nélküli kártya adatok.

Az adatforrások gyakran különböző helyekről származnak, így nem egységes formátummal jelennek meg pl. a sebesség lehet m/s, km/h de akár mérföld is. Az adatok értelmezéshez szükséges egy meta állomány, egy magyarázó, leíró adatsor, ami egyértelművé teszi, hogy mi mit jelent. Általában kijelenthető, hogy minél több adatforrást tudunk összekapcsolni, annál több új tudást nyerhetünk. Az adatforrások összekapcsolása magával hozza a redundánsan, ismétlődően megjelenő adatok kérdését, amit bonyolít, ha az adatok formátuma, mértékegysége különböző.

Az 1. ábra két közlekedéssel kapcsolatos adatforrásra mutat példát. Az első táblázatban kerékpárbérlésre vonatkozó adatok szerepelnek, de magyarázó adatok nélkül: hol van az adott számú állomás, mi a különbség a dokk és az állomás között, a kártya típuskód hordoz-e egyéb pl. korra vonatkozó információt stb.?
A második adatforrás autóbuszok által rögzített útpontok adatait mutatja. Magyarázó adatként itt hiányzik pl. az adott rendszámhoz tartozó jármű típusa, a GH, GV mezők valószínűleg koordináták, de milyen vetületi rendszerben, a sebesség ilyen alacsony vagy km/h helyett m/s-ban kell érteni.

A BigData-val kapcsolatos eljárások alapfeltétele olyan adatforrások megléte, amelyek adattartalma egyértelműen beazonosítható. Ezek az adatok nem feltétlenül közlekedési szolgáltatóktól származnak, mivel sok esetben a közösségi médiák [3], közösség által használt applikációk is szolgálhatnak adatforrásként.

3. AZ ADATFELDOLGOZÁS, ELEMZÉS FOLYAMATA

Az adatok egyik lehetséges felhasználási módja, – ami az adatbányászat egyik területe is – a csoportosítás, ill. az osztályozás. A klaszterezés során csoportokat képeznek az adatokból. A klaszterezés minősége nagyban függ az adatok jellemzőitől, ill. a hasonlóság megállapításához használt hasonlóság függvénytől. Különbséget

1. ábra: Példák adatforrások adattartalmára

	A	B	C	D	E	F	G	H
1	Kártya TAG	Mozgás időpontja	Állomás	Dokk	Mozgás típusa	Pénzmozgás	Kerékpár ID	Kártya típuskód
2	046968FA622F80	2016.09.01 4:05:41	2	3	Kibérelve (H)		G036	3313
3	046968FA622F80	2016.09.01 4:18:39	10	7	Dokkolva (K)		G036	3313
4	045B53FA622F80	2016.09.01 5:05:31	14	9	Kibérelve (H)		G114	3302
5	046027FA622F80	2016.09.01 5:05:36	15	2	Kibérelve (H)		G071	3303
6	044443FA622F80	2016.09.01 5:07:54	14	1	Kibérelve (H)		G166	3303
7	045B53FA622F80	2016.09.01 5:12:00	15	2	Dokkolva (K)		G114	3302

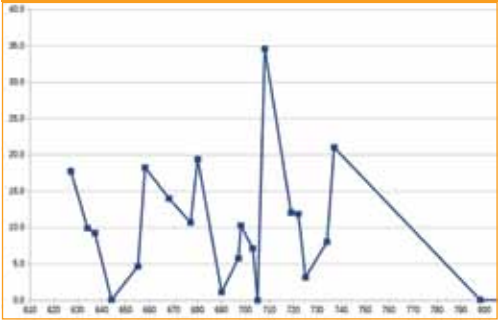
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Időpont	Rendszám	GH	GV	Sebesség	Forda	Gépkocsivezető	Járat ID	Vonalszám	Járatszám	Indulási idő	Megállóhely	Megtett m
2	2015-05-12 05:03:29.000	AFF651	17.638188	47.683653	8,272	2989-0-E	103776	1004161	GY38	1	5:04	20460	77
3	2015-05-12 05:03:34.000	AFF651	17.637550	47.683484	5,438	2989-0-E	103776	1004161	GY38	1	5:04	20460	41
4	2015-05-12 05:03:36.000	AFF651	17.637489	47.683462	4,414	2989-0-E	103776	1004161	GY38	1	5:04	20460	50
5	2015-05-12 05:04:05.000	AFF651	17.637440	47.683409	0	2989-0-E	103776	1004161	GY38	1	5:04	20460	57
6	2015-05-12 05:04:07.000	AFF651	17.637440	47.683409	0	2989-0-E	103776	1004161	GY38	1	5:04	20460	57
7	2015-05-12 05:04:23.000	AFF651	17.636553	47.683071	7,979	2989-0-E	103776	1004161	GY38	1	5:04	20460	139
8	2015-05-12 05:04:25.000	AFF651	17.636452	47.683067	7,362	2989-0-E	103776	1004161	GY38	1	5:04	20460	154
9	2015-05-12 05:04:39.000	AFF651	17.634863	47.682698	4,795	2989-0-E	103776	1004161	GY38	1	5:04	20460	263
10	2015-05-12 05:04:49.000	AFF651	17.634497	47.683129	7,614	2989-0-E	103776	1004161	GY38	1	5:04	20465	309
11	2015-05-12 05:04:51.000	AFF651	17.634300	47.683413	0	2989-0-E	103776	1004161	GY38	1	5:04	20465	11
12	2015-05-12 05:05:07.000	AFF651	17.634281	47.683436	0	2989-0-E	103776	1004161	GY38	1	5:04	20465	11
13	2015-05-12 05:05:09.000	AFF651	17.634281	47.683436	0	2989-0-E	103776	1004161	GY38	1	5:04	20465	11

kell tenni a csoportosítás és az osztályozás között. A csoportosítás vagy más néven felügyelet nélküli tanulás során, az eljárás kezdetekor nem ismert, hogy az egyes adatok, objektumok mely csoportba fognak tartozni. Az osztályozás vagy más néven felügyelt tanítás esetén az ún. osztálycímke ismert, vagyis ismert, hogy hány csoportot alakítanak ki.

Mindkét módszer alkalmazása esetén szükséges a rendelkezésre álló adatforrások hiányzó adatainak szűrése, vagy ha lehetséges a pótlása. A zajos adatok a későbbi felhasználás során problémát okozhatnak. A hiányzó adatok sok esetben pótolhatók más forrásokból, ill. a környező adatokból, amennyiben ismerjük a folyamatot, amit az adatok reprezentálnak. Az 1. ábra első táblájában események szerepelnek, amelyek egymástól függetlenek, így itt a környező adatok nem segítenek, ugyanitt a második táblán a hiányzó adat pl. megtett távolság, sebesség az előző és a következő adatból előállítható.

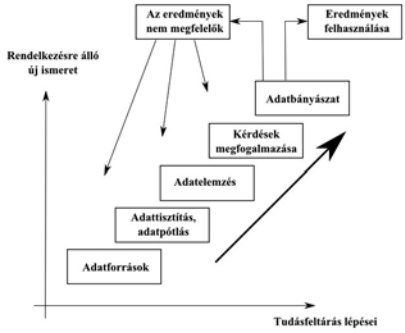
A 2. ábrán egy jármű sebessége és a megtett út diagramja látható. Kétdimenziós pontsor reprezentálja a jármű mozgását, amelyből a megtett út és a sebesség számolható. Megfigyelhető, hogy 700 m után a jármű 5 m megtétele alatt 35 k/h-s sebességre gyorsult. Jól látható, hogy itt zajos adatról van szó, amit szűrni, tisztítani szükséges a további felhasználás előtt. Lehetséges javítási megoldás az adatok legközelebbi útszakaszt reprezentáló gráfra illesztése vagy az adat eldobása. A gráfra illesztéshez szükséges egy térkép, amelynek a vetületi rendszere azonos az adatforrás vetületi rendszerével. Mint látható, ez egy újabb adatforrás.

2. ábra: Megtett út és sebesség viszonya adattisztítás előtt



A tudásfeltárás folyamatában a hibás adatok szűrése, javítása akár a teljes időráfordítás felét is elérheti. [6]. A hibás adatok ugyan gépi úton kereshetők, azonban a javítás sok esetben emberi beavatkozást igényel.

3. ábra: Tudásfeltárás folyamata



A 3. ábra a tudásfeltárás folyamatát szemlélteti. Adatforrások széles körűen rendelkezésre állnak, azonban új tudást csak akkor lehet kinyerni, ha az adatok szűrése, javítása már megtörtént. Az előzőekben bemutatott műveletek után az adatokon már végrehajthatók az adatelemzés egyes lépései. Az adatok csoportosítása, összegzése, sorba rendezése már önmagában is érdekes lehet, és iránymutatásként szolgálhat a későbbi lépésekkel kapcsolatosan. Az adatbányászatot végezhetjük irányítással, amikor van már sejtés arról, hogy mit keresünk. A másik lehetőség, amikor számítógépes eljárásokra bízzuk a folyamatot. Ekkor a vizsgálatot végző szakembernek a feladata annak eldöntése, hogy a kapott eredmény, összefüggés a számok játéka, események véletlen egybeesése vagy valódi összefüggés, milyen vizsgálatra érdemes.

4. AZ ADATOK MEGJELENÍTÉSE

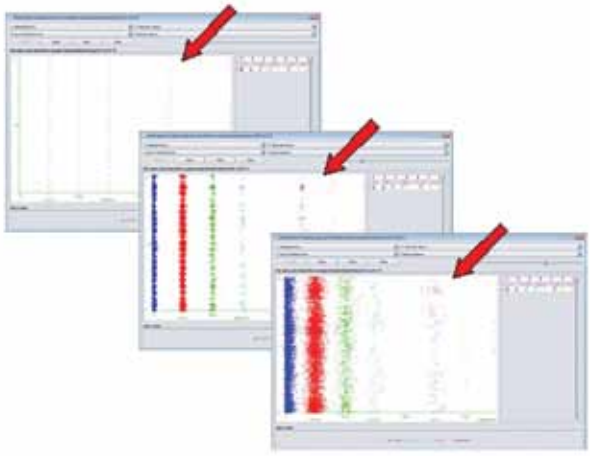
A 3. ábrán az utolsó lépés kétféle lehet, vagy a kapott eredmények felhasználása, vagy az induló feltételek módosítása, pl. az adatforrások bővítése, a tisztítás, szűrés folyamatának módosítása.

Az értékeléshez, – hogy felhasználható az eredmény vagy sem – az eredményeket meg kell jeleníteni. A táblázatos, grafikonos megjelenítésen túl rendelkezésre állnak interaktív grafikai megoldások is.

A 1. táblázat egy egyszerű adatelemzési folyamat eredményét tartalmazza. Statisztikai alapokon meg lehet állapítani annak a valószínűségét, hogy egy adott napon egy kerékpárkölcsönzőnél férfi vagy éppen nő a megjelenő ügyfél. Ez önmagában is érdekes lehet, a zöld színezés a 75% feletti, a piros a 25% alatti valószínűséget mutatja.

Ha úgy tesszük fel a kérdést, hogy marketing szempontból hol kellene a nőket és hol a férfiakat megcélozni a két táblázat érdekes eredményt mutat. A számolásban nem szerepel a nemek aránya az adott településen.

4. ábra: Tudásfeltárás folyamata



A 4. ábra a közlekedésben gyakran előforduló diszkrét értékek megjelenítésére mutat példát. Amennyiben diszkrét értéket ábrázolunk, az csak egy pontként szerepel a diagramon. Lehetséges a háromdimenziós ábrázolás azonban ebben az esetben lehet, hogy bizonyos adatok nem látszanak. Jó megoldás, ha az adatokat a pont körül véletlenszerűen szétszórjuk. A három képernyőkép három különböző állapotot mutat. Az első esetben diszkrét pontokat látunk. A második esetben egy kisebb, a harmadikban nagyobb körben szóródnak az adatok. Jól látható, hogy a második esetben a nyíllal jelzett helyen valami „furcsaság” jelent meg. Az adatokat részletesen megvizsgálva azon egy bizonyos időszakban működési probléma lépett fel. Az adatokból ez nem feltétlenül látszik, de célzott vizsgálatokkal az okok feltárhatók.

1. táblázat: Férfiak és nők megjelenési valószínűsége adott napon, adott kerékpárkölcsönzésnél

Férfi	H	K	Sz	Cs	P	Sz	V
1	21,1	22,1	24,8	17,5	15,9	15,0	14,2
2	80,4	81,3	83,3	78,5	74,2	73,0	71,8
3	56,6	58,1	61,3	50,8	47,7	46,2	44,7
4	58,6	60,0	63,3	52,9	49,7	48,3	46,7
5	73,8	75,0	77,4	69,1	66,4	65,1	63,6
6	75,8	76,9	79,2	71,3	68,6	67,4	66,0
7	66,0	67,3	70,2	60,6	57,5	56,1	54,5
8	50,2	51,7	55,1	44,4	41,3	39,9	38,4
9	56,2	57,6	60,9	50,4	47,2	45,8	44,2
10	70,4	71,6	74,3	65,4	62,5	61,1	59,6
11	85,4	86,1	87,6	82,2	80,3	79,4	78,3
12	72,5	73,7	76,3	67,7	64,9	63,5	62,1
13	50,2	51,7	55,1	44,4	41,3	39,9	38,4
14	28,4	29,7	32,6	24,0	21,7	20,8	19,7
15	47,2	48,7	52,1	41,5	38,4	37,1	35,6
16	32,0	33,3	36,4	27,2	24,7	23,7	22,5
17	50,8	52,2	55,6	45,0	41,9	40,5	39,0
18	34,5	35,9	39,1	29,5	26,9	25,8	24,6
19	34,2	35,5	38,7	29,1	26,6	25,5	24,3
20	33,5	34,9	38,0	28,6	26,1	25,0	23,8
21	49,7	51,2	54,6	44,0	40,9	39,5	38,0
22	13,8	14,5	16,3	11,2	10,0	9,5	9,0
23	23,1	24,2	26,8	19,3	17,4	16,6	15,7

Nő	H	K	Sz	Cs	P	Sz	V
1	70,9	77,9	75,4	82,5	84,2	85,0	85,6
2	19,6	18,7	16,7	23,5	25,8	27,0	28,2
3	43,4	41,9	38,7	49,2	52,3	53,8	55,3
4	41,4	40,0	36,7	47,1	50,3	51,7	53,3
5	26,2	25,0	22,6	30,5	33,6	34,9	36,4
6	24,2	23,1	20,8	28,7	31,4	32,6	34,0
7	34,0	32,7	29,8	39,4	42,5	43,9	45,5
8	49,8	48,3	44,9	55,6	58,7	60,1	61,6
9	43,8	42,4	39,1	49,6	52,8	54,2	55,8
10	29,6	28,4	25,7	34,6	37,5	38,9	40,4
11	14,6	13,9	12,4	17,8	19,7	20,6	21,7
12	27,5	26,3	23,7	32,3	35,1	36,5	37,9
13	49,8	48,3	44,9	55,6	58,7	60,1	61,6
14	71,6	70,3	67,4	76,0	78,3	79,2	80,3
15	52,8	51,3	47,9	58,5	61,6	62,9	64,4
16	68,0	66,7	63,6	72,8	75,3	76,3	77,5
17	49,2	47,8	44,4	55,0	58,1	59,5	61,0
18	65,5	64,1	60,9	70,5	73,1	74,2	75,4
19	65,8	64,5	61,3	70,9	73,4	74,5	75,7
20	66,5	65,1	62,0	71,4	73,9	75,0	76,2
21	50,3	48,8	45,4	56,0	59,1	60,5	62,0
22	86,2	85,5	83,7	88,8	90,0	90,5	91,0
23	76,9	75,8	73,2	80,7	82,6	83,4	84,2

A BigData-val kapcsolatosan a grafikus megjelenítés nagy segítség az elemzők számára, hiszen az ember számára ez sokkal többet mond, egy-egy táblázatnál.

5. A SZEMÉLYES ADATOK VÉDELME

A BigData-val kapcsolatban gyakran felmerül a kérdés, hogy hogyan biztosítható a személyes adatok védelme. Ez elsősorban az adatgazdák felelőssége, azonban a probléma elkerülhető az adatok megfelelő kezelésével. Amennyiben nem szükségesek a személyes adatok, célszerű az adatokat anonimizálni, ill. az azonos tulajdonsággal rendelkező adatokat csoportosítani, így „eltűnnek" a személyes adatok. Ez célszerűség mellett sok esetben jogszabályi kötelezettség is, azonban lehetnek olyan esetek, amikor személyes adatok szükségesek, vagy olyan adatokról van szó, amelyekből a személy beazonosítható. Megoldást jelenthet a személyes adatok helyettesítése egy kóddal, amit az adatgazda generál, így az elemzőknek ugyan személyre vonatkozó adatai vannak, de azokat nem tudja egy konkrét személlyel azonosítani. A személyes adatok és az egyéb adatok elválnak egymástól, de szükség esetén arra jogosultak az adatokat újra össze tudják kapcsolni.

6. KONKLÚZIÓ

A BigData lehetőségei, ugyanúgy ahogy más területeken a közlekedésben is óriási tartalékokkal rendelkeznek. Így a hálózattervezés, a közlekedésmenedzsment, az operatív forgalomirányítás, valamint a szolgáltatatási színvonal javításával kapcsolatos kérdésekben is nagy segítséget nyújt. Fontos azonban, hogy sok a területtel foglalkozó program önmagában nem ad eredményt. A kapott eredményeket a területhez értő szakembereknek kell értékelni és eldönteni, hogy valódi új tudásról van-e szó vagy csak a „számok furcsa játékáról" [4].

A BigData sikeres használatához három dolog szükséges, az adatforrások minél szélesebb köre, a megfelelő szaktudás és a személyes adatok tiszteletben tartása [5].

FELHASZNÁLT IRODALOM

[1] Pödör Zoltán: Az R szoftver alkalmazása az adatbányászat tárgy oktatásában Dimenziók Matematikai Közlemények III kötet, 2015 DOI: http://dx.doi.org/10.20312/dim.2015.02

[2] Big Data and Transport: Understanding and assessing options, International Transport Forum http://www.internationaltransportforum.org OECD/ITF 2015

[3] Social Media and Big Data Analysing, POSTNOTE Number 460 March 2014, The Parliamentary Office of Science and Technology, http://www.parliament.uk/post

[4] Big Data: An Overview, POSTNOTE Number 468 July 2014, The Parliamentary Office of Science and Technology, http://www.parliament.uk/post

[5] Big and Open Data in Transport New, POSTNOTE Number 472 July 2014, The Parliamentary Office of Science and Technology, http://www.parliament.uk/post

[6] Lohr, S., 2014. For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights. s.l.:New York Times.

The possibilities of using BigData in transport

Large databases and data sets are pure data cemeteries – if the data are stored and preserved – as long as they are not processed. Processing huge data sets and determining consequences poses a great challenge. As John Naisbitt said, "We are drowning in information but starved for knowledge." The tendency of recent years is to produce more and more information. According to 2015 data, the amount of data created in the world in the period 2013–2015 is the same as the total data created over the whole history preceding those years. With the evolution of technology, data registration devices and sensors have become increasingly cost effective, offering the opportunity to collect as much data as possible. The question, however, arises: why collect all this data, if these they are not used for anything later. In many cases, the data collection is not for later use, but for the support and foundation of operative procedures and decisions. It is the responsibility of engineers and data users to draw the line where the amount of data is still necessary, and to store only valuable data for the future.

Die Chancen von Big Data im Verkehrswesen

Die grossen Datenbanken und Datensätze sind reine Datenfriedhöfe, wenn die Daten nur gespeichert und aufbewahrt, aber nicht verarbeitet werden. Die Verarbeitung von grossen Datenmengen und die Gewinnung von nützlichen Zusammenhängen bedeutet eine grosse Herausforderung. John Naisbitt sagte: „Wir ersticken an Informationen, aber uns dürstet nach Wissen". Es ist der Trend der letzten Jahre, dass immer grössere Mengen von Informationen hergestellt werden. Untersuchungen von 2015 zeigen, dass die Welt im Zeitraum von 2013 bis 2015 so viele Daten erzeugt hat, wie insgesamt zuvor. Die Datenregistrierungsmittel und Sensoren sind mit der Entwicklung der Technik immer billiger geworden, und bieten sozusagen die Möglichkeit an, immer mehr Daten zu sammeln. Es erhebt sich aber die Frage: warum? – wenn diese Daten im weiteren nicht verwendet werden. In vielen Fällen werden die Daten nicht im Interesse einer weiteren Verwendung, sondern für die Unterstützung bzw. Begründung von operativen Eingriffen und Entscheidung gesammelt. Es ist die Verantwortung von den Ingenieuren und Datenbenutzern, dass sie die Grenzen zwischen den notwendigen und überflüssigen Daten ziehen.