BALOGH ZOLTÁN

Személyes adatok gyűjtése és profilírozás az interneten

Számítástudományi Tanszék

TÉMAVEZETŐ: RACSKÓ PÉTER, CSC

BUDAPESTI CORVINUS EGYETEM

Gazdaságinformatika Doktori Iskola

Személyes adatok gyűjtése és profilírozás az interneten

Doktori értekezés

Balogh Zoltán

Budapest, 2017

Ábrák jegyzéke 10

Táblázatok jegyzéke 12

1. Bevezetés 14

1.1. A kutatás célja 14

1.2. A dolgozat felépítése 17

1.3. Az alkalmazott szemléletmód és az üzleti érték 19

1.4. A szakirodalom áttekintése 21

1.5. Kitekintés 22

2. Anonimitás, adatvédelem 24

2.1 Az anonimitás definíciója 24

2.2 A személyes adatok védelme (privacy) 25

2.3 Törvényi szabályozás 27

2.4 Az anonimitás keresésének okai 28

2.5 Az anonimitás elhagyásának következményei 31

3. A látogatók azonosításának és követésének módszerei 33

3.1. A kliensoldal 33

3.1.1. Böngészők 33

3.1.2. A munkamenet 37

3.1.3. A kliens oldali adattárolás 38

3.2. A szerveroldal 40

3.3. A böngészés folyamata 42

3.3.1. A weblap megtekintése 42

3.3.2. A böngészés révén kinyerhető adatok 42

3.3.3. A rendelkezésre álló adatok 45

3.3.4. A hozzáférhető adatok 46

3.3.5. A felhasználói életút vizsgálata 51

3.4. Látogatók azonosításának és követésének módszerei 53
3.4.1. A böngésző által szolgáltatott adatokból kinyert információ segítségével 54
3.4.2. A viselkedés alapú nyomkövetés 59
3.4.3. A látogató azonosítása és követése 61
3.5. Összegzés 63
4. Az eszközböngészőkről begyűjthető adatok jellemzői 65
4.1. A kutatás bemutatása 65
4.2. Az adatgyűjtés 65
4.2.1. A célcsoport 68
4.2.2. Az adatgyűjtő alkalmazás 68
4.2.3. A lementett adatok 70
4.3. A bizonytalanság-csökkentő képesség mérőszáma 71
4.4. Az összegyűjtött adatok vizsgálata 72
4.4.1. Alapvető statisztikák 73
4.4.2. A lementett változók bizonytalanság csökkentő ereje 75
4.5. Összegzés 77
5. Az Egyetemi polgárok pszichológiai jellemzői adatvédelmi szempontból 79
5.1. A myPersonality Project 80
5.2. A Big 5 – személyiségi jellemzők 81
5.3. Apply Magic Sauce 82
5.4. Az adatelemzés 83
5.4.1. A myPersonality Project adatai 84
5.4.2. A statisztikai elemzés 84
5.4.3. Klaszterelemzés 87
5.4.4. Az eredmények kiértékelése 95
5.5. Összegzés 96

6. Személyes információ kinyerése webes adatokból 98
6.1. Az adatok előkészítése 99
6.2. Elemzés 100
6.2.1. Klasszifikációs algoritmusok 100
6.2.2. Asszociáció 103
6.3. Összegzés 105
7. Utóhang 107
8. Függelék 110
8.1. HTML 110
8.2. HTTP lekérdezés 110
8.3. Javascript 114
8.4. Süti (cookie) 115
8.5. CSS3 és HTML5 képességek 118
8.6. Látogatók azonosításához használható kiemelt paraméterek elemzése 119
8.7. Adatgyűjtő alkalmazás 124
8.7.1. Az alkalmazás működése 125
8.7.2. Hozzáférhető adatok relevanciája és terhelése 127
8.7.3. A lementett adatok listája 132
8.7.4. A lementett adatok statisztikai jellemzői 136
8.7.5. A feldolgozott adatok listája 138
8.8. Idézett források 141
8.8.1. Tudományos szakirodalmi művek 141
8.8.2. Hírek és cikkek 146

Ábrák jegyzéke

1. ábra: Filter bubble grafikus ábrázolása (Pariser, 2011) 31
2. ábra, az index.hu HTTP lekérdezés fejléce (saját felvétel) 36
3. ábra, A GMail nem működik sütik használata nélkül (saját szerkesztés) 38
4. ábra, követésre alkalmas technológiák fejlődése (saját) 40
5. ábra: webszerver kapcsolata az adatbázissal és a böngészővel (Bhavin, 2014) ... 41
6. ábra, HTTP lekérdezés fejléce (saját felvétel) 41
7. ábra, Egy harmadik fél által írt alkalmazás engedélyeket kér a Facebook-tól (saját felvétel) 45
8. ábra, a látogató és az Internet közötti kapcsolat (saját szerkesztés) 46
9. ábra, a Chrome engedélyt kér a látogató földrajzi pozíciójához (saját felvétel) ... 47
10. ábra, asztali operációs rendszerek piaci részesedése 2016-ban (NetMarketShare, 2016) 50
11. ábra, Egy webhely weblapjainak gráf formában történő ábrázolása (saját szerkesztés) 51
12. ábra, Azonosítás módszereinek egymáshoz való viszonya (saját szerkesztés) 53
13. ábra, Cache alapú azonosítás 58
14. ábra: a földrajzi pozíciók város, utca és házszámra történő feloldása Pentaho-val Google Maps API-n keresztül 71
15. táblázat: Lementett változók bizonytalanság csökkentő ereje 76
16. táblázat: Lementett változók bizonytalanság csökkentő ereje 77
17. ábra: Apply Magic Sauce - Prediction API alcíme (The Psychometrics Centre, 2013) 82
18. ábra: Facebook Like-okat személyiségi jegyekké feloldó Pentaho alkalmazás (saját szerkesztés) 84
19. ábra: Az intelligencia és az élettel való elégedettségét megjelenítő scatterplot diagram 86
20. ábra: A klaszterek 3 dimenziós hisztogramja 89
21. ábra: Hierarchikus klaszterelemzés dendrogramja 94
22. ábra, URL szerkezete (saját szerkesztés) 110
23. ábra, A HTTP lekérdezés menete (Websiteoptimization.com, 2009) 111
24. ábra, a WizzAir.hu főoldalának HTTP válasza (saját) 113
25. ábra, HTML DOM objektum kezelése Javascript-tel (saját szerkesztés) 115

26. ábra, szerver által küldött sütik (saját felvétel) 116
27. ábra, süti létrehozása Javascripttel és jQuery segítségével (saját) 116
28. ábra: Adatgyűjtő alkalmazás megjelenése a Corvinus e-learning rendszerében (saját felvétel) 125
29. ábra: Adatgyűjtő alkalmazás vázlatos működése (saját szerkesztés) 127
30. ábra: Azonosításra használható eszközböngészőből kinyerhető paraméter és felvehető értékei (saját szerkesztés) 135

Táblázatok jegyzéke

1. táblázat: A Facebook különféle típusú hirdetéseinek átkattintási rátája 22
2. táblázat: 5 legelterjedtebb böngésző (W3Schools, 2017) 34
3. táblázat: Néhány ismertebb böngésző és megjelenítő motorja (Stanclift, 2008)... 35
4. táblázat: harmadik féltől lekérdezhető paraméterek előnyei és hátránya 48
5. táblázat: azonosításra használható paraméterek és azonosításuk tárgya 55
6. táblázat: Apply Magic Sauce API-ja által visszaadott személyes adatok (The Psychometrics Centre, 2013) 83
7. táblázat: A Big 5 jellemzőinek átlaga és 95%-os konfidencia intervalluma 85
8. táblázat: A Big 5 jellemzőinek átlaga és 95%-os konfidencia intervalluma 85
9. táblázat: A Big 5 jellemzőinek átlaga és 95%-os konfidencia intervalluma 85
10. táblázat: A Big 5 jellemzőinek átlaga és 95%-os konfidencia intervalluma 86
11. táblázat: A myPersonality Big 5 jellemzőinek ANOVA táblája 87
12. táblázat: A myPersonality politikai jellemzőinek ANOVA táblája 88
13. táblázat: A myPersonality családi állapot jellemzőinek ANOVA táblája 88
14. táblázat: A klaszter tagság kereszttáblája 88
15. táblázat: A Big 5 és a családi állapot klasztereinek ANOVA táblája 89
16. táblázat: Végső klaszter középpontok 90
17. táblázat: Klaszterek kereszttáblája 90
18. táblázat: Big 5 klaszterek ANOVA táblája 91
19. táblázat: Politikai nézet ANOVA táblája 91
20. táblázat: Családi állapot ANOVA táblája 92
21. táblázat: Családi állapot kereszttábla 92
22. táblázat: A BIG5 és a családi állapot változóinak ANOVA táblája 93
23. táblázat: Klaszter tagság ellenőrzés 95
24. táblázat: myPersonality Project adatainak K-középpontú klaszterei 95
25. táblázat: a saját adatok K-középpontú klaszterei 95
26. táblázat: az alkalmazott osztályozó algoritmusok által helyesen osztályozott változók százalékos aránya 102
27. táblázat: az intelligencia és az élettel való megelégedettség változókra alkalmazott klasszifikációs algoritmusok kimenete 103
28. táblázat: a talált szabályok bizonyossági (confidence) szintje rögzített bizonyosság és támogatottsági szint mellett 104

29. táblázat: a talált szabályok bizonyossági (confidence) szintje rögzített bizonyosság és támogatottsági szint mellett 105
30. táblázat: A GET és POST lekérdezés HTTP fejléce (saját) 113
31. táblázat: az SQL és a HTTP CRUD műveletei (Medic, 2014) 113
32. táblázat: kiemelt paraméterek elemzése 121
33. táblázat: a lementett paraméterek statisztikai jellemzői 137

1. BEVEZETÉS

2013. március 12-én lett 25 éves az internet. Valószínűleg a HTTP protokoll és a HTML nyelv feltalálója Tim Berners-Lee sem gondolta akkoriban (Owen, 2014), hogy forradalmasítani fogja az emberi kommunikációt. Ő maga 2013. március 18-án megkapta az Erzsébet Királynő Mérnöki Díjat (Queen Elizabeth Prize for Engineering) találmányáért. A statisztikák szerint 2012-ben az emberiség több mint egyharmada használta az Internetet és 2016-re az éves forgalmazott adat mennyisége elérte az 1 zettabyte-ot is (Cisco, 2014). 2014-ben hozzávetőlegesen 600 millió weblap létezett, melyek végérvényesen megváltoztatták az emberi információszerzés és tartalommegosztás módját.

A web fejlődés egyik mérföldkövének a közösségi oldalak megjelenését tekinthetjük, a legjelentősebbnek számító közösségi háló, a Facebook megreformálta az információ elérésének módját. (Jeffrey, 2012) 2017 februárjában a közösségi hálónak több, mint 1 milliárd 860 millió felhasználója és több, mint 1 milliárd 150 millió látogatója volt naponta. (Zephora - Digital Marketing, 2017)

1.1. A kutatás célja

Egy 2015-ös felmérés szerint a britek többsége a mindenki számára hozzáférhető, cenzúramentes és anonim internet mellett teszi le a voksát. (Healy, 2015) Ahogy az internet a hétköznapjaink részévé vált, a weboldalakat megjelenítő alkalmazások – a böngészők – fejlődése is felgyorsult. Ez lehetővé tette a weboldalak számára, hogy egyre több adatot érjenek el a látogatóikról. Az online tartalmak megtekintésével megismerhetővé vált a látogatók fogyasztóinak preferenciái és böngészési szokásai. Ezen adatok elemzésével a látogató személyre szabott tartalmat kaphat és ezzel egyidejűleg felfedhetővé válnak a személyes jellemzői is. A személyre szabott tartalmak következtében megjelent a „filter bubbles” jelensége, amely során a felhasználó az észlelt tulajdonságaihoz illő tartalmakhoz fér hozzá, viszont nincs kontrollja afelett, hogy mi az, amit fogyaszthat és mi az, amit nem. A felhasználók profilozása miatt a kezdeti anonim web ma már nem az.

A látogatók preferenciáinak megismerését követően megjelentek az online ajánlórendszerek. Például, az amerikai online műsorszóró vállalat, a Netflix ajánlórendszere hozzávetőlegesen az esetek 80%-ban befolyásolja a látogatókat a tartalmak fogyasztásában. (Carlos & Neil, 2015) Emiatt nem meglepő, hogy a cég

igencsak érdekelt az általuk használt algoritmusok hatékonyságának növelésében. Emiatt 2006 és 2009 között 1 millió dolláros Netflix Díjjal ösztönözték vállalkozó szellemű vállalkozókat az ajánlórendszer hatékonyságának javítására. A győztes csapat által kidolgozott algoritmus hatékonyságában 10,06%-kal múlta felül a Netflix által használtat. (Lohr, 2009) Az internetes szolgáltatók komoly erőfeszítéseket tesznek ügyfeleik profiljának minél pontosabb meghatározására, mert ez számukra versenyelőnyt, sokszor pedig a pályán maradás feltételét jelenti. Nyugodtan kimondhatjuk, hogy ami az interneten megismerhető, azt a szolgáltatók meg is fogják ismerni. A dolgozatban azt kutattam, hogy melyek az interneten egy-egy felhasználóról legális eszközökkel kibányászható adatok és az adatokból levonható következtetések határai.

Az információszerzés és felhasználás területén jelenleg paradigmaváltás figyelhető meg: amíg korábban a felhasználók pusztán a korszerű keresőmotorok képességeit kihasználva találták meg a számukra releváns információt, mára a keresőmotor használatán felül az online szolgáltatások a látogatók preferenciájának megismerésével képesek eldönteni, hogy egy adott tartalom a látogató számára érdekes lehet-e, azaz a tartalom jellemzőit és a látogató preferenciáit összevető algoritmus dönti el, hogy az a látogató számára érdekes-e. A preferenciákat összeállító, profilozó algoritmust használó oldalak (keresőoldalak, hírportálok és közösségi oldalak) használatával a látogatók implicit módon hozzájárulnak a detektálható jellemzőik gyűjtéséhez, majd ezen adatok alapján az oldal tartalmakat ajánl számukra fogyasztásra. A weboldalak által használt ajánlórendszerekről a látogatók sok esetben nem tudnak és használatuk mellőzésére az esetek többségében nincs mód.

Kutatásaim során arra kerestem a választ, hogy a weboldalak számára mely hozzáférhető adatokból lehetséges a látogató tulajdonságaira következtetni, mely adatok alkalmasak a profilok összeállítására és ez hogyan történik. Kutatásaimat az alábbiak szerint építem fel:

Az első kutatási cél a nem professzionális, **egy domain alól elérhető weboldalak[1] vizsgálata**. Exploratív kutatás keretében az internetezéshez használt

[1] Általában az egy domain alatt elérhető weboldalak képesek hozzáférni a böngészéshez használt eszköz, a rajta lévő operációs rendszer és a böngésző valamennyi tulajdonságához (céges weboldak, hírportálok, blogok, webáruházak), feltétel, hogy nincs az oldalnak más weboldalakba beépülő adatgyűjtő modulja, amellyel a látogatók preferenciáit vagy böngészési jellemzőit lehetséges feltérképezni, a látogatók nem regisztrálják magukat az oldalra, amely esetben a beazonosítás triviálissá válna

böngészőkből, valamint az internetezéshez használt hardverről és annak szoftver-környezetéből kinyerhető adatokat elemezve következtettem a felhasználó személyes jellemzőire. Az elemzési fázisban a felhasználók és személyes paramétereik között csak triviális kapcsolatot sikerült kimutatni, a magukat szándékosan felfedni nem kívánó látogatók esetében nem sikerült személyes jellemzőket megállapítani. Megvizsgáltam a böngészők számára **hozzáférhető paraméterek bizonytalanság-csökkentő erejét** is. Felmértem az egy domain alól elérhető weboldalak és közösségi oldalak látogatóikról elérhető adatok mennyiségét és minőségét. A látogatókról összegyűjtött adatok bizonytalanság-csökkentő képessége megmutatja, hogy a mintán belül mekkora valószínűséggel található meg egy egyed, és összehasonlíthatóvá válik a kinyert paraméterek különböző csoportosításainak információhordozó ereje.

A második kutatási cél a Budapesti Corvinus Egyetem **közösségi oldalak** által összegyűjtött, **kinyilvánított preferenciákból** kinyert **személyes tulajdonságok** alapján a látogatók csoportosítása, majd a kapott csoportok összevetése a myPersonality Project (Stillwell & Kosinski, 2012) során összegyűjtött adatokból készített klaszterekkel. Azt mutatom be, hogy a felhasználók által önként szolgáltatott adatokból hogyan lehet személyiségre vonatkozó következtetéseket levonni. Természetesen az adatokat felhasználás előtt anonimizáltam. A kutatás során a Facebook-tól letöltött egyénekhez köthető „Like" adatbázist elemeztem.

A közösségi hálózatok aktív tagjai az online tartalmakról alkotott tetszésüket a „Like" [2] gombra történő kattintással is kifejezhetik. Ez a látogatókhoz köthető információ az közösségi hálózatokon elérhető. A kutatás során a Budapesti Corvinus Egyetem polgárainak „Facebook Like"-jait elemeztem pszichológiai API segítségével (Kielczewski, 2017), majd az egyéneket a kapott személyes jellemzőik alapján nem felügyelt tanulási módszerekkel klaszteresztem. A kutatás szintén feltáró jellegű, a Budapesti Corvinus Egyetem polgárainak és a myPersonality Project résztvevőinek a pszichológiai API által visszaadott személyes tulajdonságokból képzett látogatói klaszterek közötti különbségeket mutatja be.

A harmadik kutatási célom annak kikísérletezése volt, hogy hogyan lehet **a látogatók nem személyes jellemzőiből következtetni személyes jellemzőikre**. Az Apriori algoritmus (Gautam, Ghodasara, & Parsania, 2014) használatával a Budapesti

[2] Facebook Like gomb (2010 második negyedév): a felhasználók kifejezhetik a tetszésüket egy weben található tartalom iránt. Ezzel a lépéssel azokról a weboldalakról is képes a Facebook adatokat gyűjteni a felhasználóiról a közösségi oldal meglátogatása nélkül

Corvinus Egyetem polgárainak e-learning környezetbeli viselkedését elemezve arra kerestem a választ, hogy a látogatók mely személyes tulajdonságaikra lehetséges online viselkedésükből következtetni.

A dolgozat további, a kutatás technikai részét tartalmazó részében bemutatom a különböző típusú weblapok által hozzáférhető adatokat és a belőlük kinyerhető, a látogatókra vonatkozó személyes jellemzőket. Amíg az egy domain alól elérhető weboldalak a látogatókról csekély személyes információt képesek kinyerni, egy kiterjedt, beépülő modulokkal rendelkező oldal[3] átfogó profilt képes építeni a látogatói preferenciáiról, érdeklődési körükről és szokásairól. (Szommer, Balogh, & Racskó, 2014).

A látogatók személyes tulajdonságuk alapján történő csoportokba rendezése üzletileg jól hasznosítható eredményt hoz, ui. az egyes csoportoknak célzott reklámok küldhetőek. A kezdetek óta a Facebook egyik üzleti stratégiai alappillére a közösségi hálózatokban rejlő hirdetési felület adta lehetőség kiaknázása. (Jeffrey, 2012) A weboldalak látogatói az üzlet szempontjából meghatározó tulajdonságaik alapján csoportosíthatóak.

Az üzleti szempontból megfelelő minőségű felhasználói profilok kialakításához elengedhetetlen a látogatókról gyűjtött adatok időközönkénti frissítése. Emiatt és a felhasználók megtartása érdekében a Facebook folyamatosan fejleszti termékeit[4] és azon dolgozik, hogy a felhasználóiról egyre több és pontosabb információt szerezzen be[5][6]. (Zuckerberg, 2015).

1.2. A dolgozat felépítése

A dolgozatom központi témáját – a kibertér szereplőiről elérhető információk minőségi jellemzését – három saját kutatás segítségével mutatom be a következő bekezdésekben bemutatott vezérfonal mentén.

---

[3] A kiterjedt hálózattal rendelkező weblapok csoportja (közösségi hálózatok, hirdetési ügynökök stb.) olyan weboldalakat foglal magában, amelyek vagy a moduljait más webloldalakba beépítették és amelyek képesek az oldalt látogatókról adatokat küldeni a beépülő modulok gazdáinak, amelyek révén lehetséges a látogatók preferenciáinak feltérképezése.

[4] Facebook Home (2013 első negyedév) alkalmazás/operációs rendszer kiegészítő modul megjelenése Androidra, melynek segítségével egy átlagos weblap vagy okostelefon alkalmazáshoz képes jóval több mélyebb szinten ágyazódik lesz része a mobil operációs rendszernek és több adathoz is hozzáfér a felhasználóról

[5] Egérmozgás figyelés (2013 október): bizonyos felhasználók csoportjának egérmozgásának figyelése (Rosenbush, 2013) amelyből kinyert adatok segítségével értéknövelt szolgáltatás nyújtása

[6] Fejlődő országok területén Internet szolgáltatása drónok segítségével (2015. március) Mark Zuckerberg bejelentette, hogy a fejlődő országokba drónok segítségével fogja az Internetet szolgáltatni (Camilla, 2015)

Az empirikus, irodalmakat áttekintő részben az online anonimitás témakörének fogalmait definiálom, elemzem az elhagyásának következményeit, végül meghatározom a dolgozatban alkalmazott személeletmódomat. (Xu, Dinev, & Smith, 2011)

A technikai bevezető részben a látogatók azonosításának és követésének módszereit a webes böngészés alapjaitól kezdve strukturált formában foglalom össze, ezt követően a látogatókról a böngészőn keresztül kinyerhető tulajdonságokat rendszerezve közlöm, majd azok hozzáférhetőségének módját vizsgálom.

Az elméleti felvezetés után következik a saját kutatásaimat leíró rész, amelyekben a magukat felfedő és a magukat felfedni nem kívánó látogatókat egyaránt vizsgálom. Mindhárom kutatás közös jellemzője az közös adatgyűjtés, amelynek részletes leírása a *4.2 Az* adatgyűjtés részben olvasható.

A látogatók beazonosíthatóságának elemzése igen aktuális téma. Rengeteg publikáció született a témában, amelyeknek jelentős részének szerzői a HTTP szerver log-okban kutatva keresi a látogatóhoz tartozó munkameneteket minták után kutat. A Myriam Abramson és David W. Aha User Authenticaion from Browsing Behaviour című publikációjában leírt algoritmus a szerver log-okban keresi az egyes felhasználóra jellemző mintákat. (Abramson & Aha, 2013). Dominik Herrmann, Christoph Gerber, Christian Banse és Hanness Federrath „Analyzing Characteristic Host Access Patterns for Re-Identification of Web User Sessions" írásában Bayes osztályozó algoritmus segítségével keresi a különböző felhasználókhoz tartozó munkameneteket, hasonlóan az Iváncsy Renáta és Juhász Sándor által írt „Analysis of Web User identification Methods" műben elemzett cookie nyomkövetés elemzéshez. (Iváncsy & Juhász, 2007) A *4. Az eszközböngészőkről*[7] *begyűjthető adatok jellemzői* fejezetben leírt szintén a weboldalak számára hozzáférhető adatokból táplálkozva igyekeztem a felhasználókat beazonosítani és tulajdonságaikra következtetni, valamint a felhasznált paraméterek bizonytalanság-csökkentő erejét vizsgáltam.

A Budapesti Corvinus Egyetemi polgárok pszichológiai jellemzőinek átlagtól való eltérései című kutatás során az egyetemi polgárok összegyűjtött Facebook Like-jait elemeztettem a myPersonality Project pszichológiai API-jával, majd a kapott

[7] Az eszközböngésző egy általam bevezetett fogalom egy konkrét böngésző alkalmazás egy példánya, amelyet az egyén a weboldal eléréséhez használt egy internetképes eszközén. Ezáltal lehetséges a különböző internetezésre alkalmas eszközökön található böngésző fajta és egy konkrét eszközön található konkrét böngésző megkülönböztetése.

eredményeket összevetettem más országok eredményeivel. A kutatás kiegészíti Dr. Michal Kosinski és Dr. David Stillwell myPersonality Project-tel kapcsolatos kutatásait.

Az *Egyetemi polgárok pszichológiai jellemzői* című kutatásban a látogatók nem személyes jellemzőiből következtetek a személyes jellemzőikre.

A kutatásaim során átfogó képet mutatok be a különböző méretű és kiterjedésű weboldalak látogatói adataihoz való hozzáférésről, valamint ezen adatokból kinyerhető információ minőségéről. Az eszközböngészőkről begyűjthető adatok fejezetben felmérem az egy domain alól elérhető weboldalak képességeit, a Budapesti Corvinus Egyetem polgárainak pszichológiai jellemzői fejezetben a kiterjedt hálózattal rendelkező oldalak (pl: közösségi hálózatok) képességeit elemzem.

1.3. Az alkalmazott szemléletmód és az üzleti érték

A webet használó közösség tagjai előszeretettel használják a mindenki számára hozzáférhető „ingyenes" online szolgáltatásokat: „ingyenes" tartalmakat fogyaszthatnak, kommunikálhatnak ismerőseikkel vagy tartalmakat oszthatnak meg. Ezen szolgáltatások használatával a felhasználók önként fedik fel sokszor felelőtlenül magánéletük minden egyes mozzanatát, de elvárják az online szolgáltatásoktól, hogy védve legyenek a rosszakarók ellen. Az üzleti érdekek azonban az online anonimitás felszámolását sürgetik, emiatt az elmúlt néhány a jelentősebb online informatikai szolgáltatók sorra jelentették be, hogy nem támogatják többé az anonim felhasználókat.

Mark Zuckerberg a Facebook alapítójának testvére Randi Zuckerberg, a Facebook marketing igazgatójának nyilatkozata szerint az anonim felhasználókból nem lehet pénzt csinálni (Chen, 2011) és a jövőben folytatni fogják a harcot az online anonimitás felszámolásáért. A közösségi oldalak ilyen mértékű terjedésének kétségkívül vannak hátrányai is: külön-külön többet tudnak rólunk, a felhasználókról, mint az FBI és a CIA együttvéve.

2012. januárjában a Google bejelentette az új adatvédelmi irányelveit, amelynek értelmében a Google által birtokolt online szolgáltatások ezt követően megoszthatják a felhasználókról gyűjtött adatokat egymás között.[8] Ezzel a lépéssel a

[8] Ez a gyakorlatban azt jelentheti, hogy a délután egykor Budapesten tartózkodó személyt figyelmeztetheti a Google, hogy induljon el a délután 16:00-kor kezdődő debreceni megbeszélésére egy fél órával korábban, mert az M3-as autópályán baleset történt, amennyiben a találkozót felvette a Google Naptárába.

Google emelt szintű szolgáltatást képes nyújtani a felhasználói számára, de többet is fog tudni a használói cselekedeteiről és szokásairól.

„It will be very hard for people to watch or consume something that has not in some sense been tailored for them."

Eric Schmidt, Google

Az anonimitás témakörében az informatikával foglalkozó óriások állásfoglalása mára többé-kevésbé eldőlt. Az utóbbi években egymást követően jelentették ki, hogy a jövő igenis a személyre szabott tartalmaké lesz. Ez pedig azt vonja maga után, hogy az általuk üzemeltetett oldalakon szükséges a látogatók azonosítása és követése. Ezáltal személyre szabott tartalmakat és reklámokat küldhetnek számukra.

„A squirrel dying in front or your house may be more relevant to your interests right now than people dying in Africa."

Mark Zuckerberg, Facebook

Az üzleti érdekek következtében tehát nem az anonimitásra való törekvés a fő elv a weboldalak kialakítása során, hanem az, hogy a felhasználókról minél több információt érjenek el. Emiatt egyre jelentősebb különbség van az közösségi hálók felhasználói által szándékosan közzétett adatok és a róluk ténylegesen az interneten elérhető adatok között.

Egyes vélemények szerint az anonim kommunikáció jelenléte veszélyezteti a társadalom alappillérét: az elszámoltathatóságot és a felelősségre vonhatóságot. (Davenport, 2002)

Évente egyre több időt töltünk az online világban, emiatt a nekünk szóló reklámok is követnek minket oda. Az online világban feladott hirdetések abban különböznek a valós életbeli, hogy a hirdetés feladójának bővebb lehetősége van információt szerezni a hirdetést megtekintő személyéről. A hirdetők célja az, hogy minél több és relevánsabb információkat érjenek el a hirdetés megtekintőjével kapcsolatban. Minél több adat elérhető a látogatókról, annál pontosabb képet kapnak róluk, feltárulnak a szokásaik, vágyaik és ízlésük.

Kutatásaim során nekem is az a célom, hogy a látogatókról minél több és értékesebb információt nyerjek ki. A dolgozatom hangvétele magán hordozza a tanulmányaim során felvett gazdaságinformatikusi szemléletmódot, amelyet Cser László az alábbi módon definiált:

„A gazdaságinformatika a közgazdasági és az informatikai tudományok ismereteinek egyfajta kombinációja. A fogalom az üzleti szférában és a gazdálkodási területeken kezelt szociotechnikai rendszereket, az emberek és gépek által fejlesztett, illetve kezelt információ- és kommunikációrendszereket jelenti. A középpontban a gazdasági/üzleti feladatok támogatása áll." (Cser, Nagyné Polyák, & Németh, Informatikai Alapok, 2007)

1.4. A szakirodalom áttekintése

„The 21st century... when deleting history is more important than making it...?"
(Smart, 2013)

A feldolgozott irodalmat az irodalomjegyzékben először a tudományos jellegük szerint csoportosítva, ezen belül pedig szerző szerinti ABC sorrendben közlöm, az alábbi kategóriákba sorolva:

- **tudományos szakirodalomi művek**: szakkönyvek, konferenciaközlönyökben vagy kiadványokban megjelent publikációk, cikkek
- **hírek és cikkek:** nem a nyomtatott sajtóban megjelent cikkek, közlemények vagy leírások

Az elméleti részek témája az online anonimitás és a privacy fogalmához kapcsolódik, a gyakorlati részhez pedig az adatgyűjtés és az adatok feldolgozása során használt technológiák és algoritmusok szakirodalma tartozik. A feldolgozott irodalmak tehát az alábbi kategóriák valamelyikébe sorolhatók be:

- Anonimitás/privacy fogalmát
    - A törvényi hátterét érintő
- Közösségi hálózatokról szóló
- Webes technológiákat taglaló
- Adatbányászati algoritmusokkal foglalkozó

Az anonimitással foglalkozó irodalom feldolgozása során a definícióból kiindulva fejtegetem az anonimitás fenntartásának okait. A téma tárgyalása alkalmával kitérek a magyar, európai és az amerikai törvényi háttérre. Ezt követően áttekintem a különféle embertípusok anonimitáshoz való kapcsolatát és az anonimitás elhagyásának következményeit.

A webes technológiákat taglaló irodalom főként a *3. A látogatók azonosításának és követésének* módszerei fejezetben jelenik meg, hiszen ez a fejezet foglalkozik a téma technológiai részével, az adatbányászati algoritmusokkal pedig a *6. Személyes információ kinyerése webes adatokból* fejezet.

1.5. Kitekintés

A bevezetés utolsó alfejezetében néhány – többségük a Facebook-kal kapcsolatos – az internetes anonimitás jövőjét befolyásoló cikket emelek ki.

Látva a Facebook sikerét az elmúlt években sok közösségi oldal látott napvilágot, amelyeknek alapvetően nem az a célja, hogy a Facebook felhasználók zömét magukhoz csábítsák, hanem a már jelenlévő sikeres közösségi oldalak pozitívumait és az újonnan felmerült felhasználói igényeket szem előtt tartva fedjenek le egy-egy részpiacot. Az új beszállók célja, hogy kihasítsanak maguknak egy szeletet az online reklámpiaci tortából.

Az iparági átlagos átkattintási ráta (CTR[9]) az összes hirdetési formátumra 0,06% (Chaffey, 2015), ehhez képest a Facebook átkattintási rátája 2013-ban hirdetéstől függően 0,02%-3,20% is lehet (Salesforce, 2013).

Hirdetés típusa	**CTR**
Külső weboldalon lévő hirdetés	0,02%
Alkalmazás	0,04%
A felhasználók idővonalán megjelenő írás	2,03%
Szponzorált helyről történő felhasználói bejelentkezés	3,20%

1. táblázat: A Facebook különféle típusú hirdetéseinek átkattintási rátája

Ennek ellenére néhány statisztika szerint a fiatal korosztályban a Facebook már nem számít menőnek, mivel azon már a szüleik és nagyszüleik is regisztráltak. (Christina, 2013) Emiatt a Facebook elkezdte felvásárolni azokat az online szolgáltatásokat, amelyeket a fiatalok előszeretettel használnak, például az Instagram és a Whatsapp. Valamint 2015. márciusában Mark Zuckerberg bejelentette, hogy már tesztelik az internetet szórni képes drónokat, amelyeket a fejlődő országokban szeretnének bevetni a közeljövőben. (Camilla, 2015)

[9] click through rate - átkattintási ráta: kattintásra jutó megtekintések száma

A trendek abba az irányba mutatnak, hogy a webes technológiák – HTML5 és Javascript API[10]-k – fejlődésével az online alkalmazások számára egyre több adatot képesek elérni a látogatók által használt eszközökről. Az összegyűjtött felhasználóhoz rendelt adatok a közösségi oldalak API-jainak használatával harmadik fél számára kinyerhető a felhasználók hozzájárulásával.

A közösségi hálózatok piacán kiélezett verseny zajlik a felhasználókért, a Facebook igyekszik a jelenlegi 1,8 milliárdos felhasználóbázisát még tovább duzzasztani, valamint a világ jelentős részén[11] az egyeduralmát hosszútávon bebiztosítani, amihez a HTML nyelv és a futtató platformok fejlődése nagyban hozzásegít, mivel rajtuk keresztül egyre több adatot hozzáférhető a felhasználókról.

[10] Application Programmers Interface: egy program azon eljárásainak (szolgáltatásainak) és azok használatának dokumentációja, amelyet más programok felhasználhatnak

[11] Az alábbi országokban az alábbi közösségi hálózatok a legnépszerűbbek: Kínában a QZone; Oroszországban, Kazahsztánban, Belorussziában, Ukrajnában, Fehéroroszországban a VK (VKontakte); Iránban a Facenama; Japánban a Twitter (Vincenzo, 2016)

2. ANONIMITÁS, ADATVÉDELEM

Elsőként a releváns irodalmat ismertetem, amelyek az anonimitás fogalmával, a jelenlegi személyes adatainkat védő szabályozással, valamint az anonimitás keresésének és elvesztésének következményeivel foglalkoznak.

2.1 Az anonimitás definíciója

Ahogyan a valós életben, az emberek online is sok esetben burkolózhatnak anonimitásba (Rigby, 1995). Ilyen esetek lehetnek például, ha egy online fórumon valamilyen kényes témáról beszélgetnek (betegségek, kisebbségi kérdések, szexuális élet stb.) és a beszélgetés alanyai nem szeretnék felfedni valódi kilétüket, mert attól tartanak, hogy a valós életben az online tett kijelentésük miatt megbélyegeznék őket.

Gary Marx (Marx, 1999) definíciója szerint az anonimitás azt jelenti, hogy az általa meghatározott 7 azonosítási dimenzió egyik tulajdonsága sem ismert az azonosítandó illetőről. Az azonosítás 7 dimenziója:

- név
- helyzet
- álnév, amely utal a személy nevére vagy helyzetére
- álnév vagy becenév, amely egyértelműen nem utal a személy nevére vagy helyzetére, de a kérdéses személy nyomára vezethet
- viselkedési minták
- társadalmi réteghez való tartozás
- információ vagy tudás, amely jellemző a személyre

A listában található elemek egy része teljesen konkrétan beazonosítja az adott személyt, a másik része olyan tudásra utal, amely segítségével kikövetkeztethető az egyén személyazonossága. Tehát az anonimitást kereső személyek elkerülik a fentebb említett 7 kategória valamelyikébe sorolható személyes adataikat felfedjék. A felsorolásban a viselkedési minta kulcsfontosságú, hiszen a kutatásaim során a látogatókat az online környezetben kinyilvánított viselkedésük és preferenciáik alapján azonosítom be.

Andreas Pfitzmann és Marti Hansen szerint az anonimitás egy olyan állapot, amelyben nem lehetséges egyértelműen beazonosítani egy tett végrehajtóját. A

potenciális elkövetők halmazát az **anonimitás készletnek** (anonymity set) nevezi, amelyek közül bármi egyenlő eséllyel lehet az elkövető. (Andreas & Marit, 2010)

Bizonyos esetekben, például pénzügyi tranzakciók esetében az egyének szintén anonimitásba burkolózhatnak, de a későbbiekben felfedhetik a kilétüket. Az egyén kiléte felfedését követően már nincs további lehetősége újból anonimitásba burkolózni.

2.2 A személyes adatok védelme (privacy)

Amíg az anonimitás az elkövető személyének felfedhetetlenségével, a személyes adatok védelme (privacy) pedig a (bűn)elkövető által elkövetett cselekmények elrejtésével foglalkozik. (Mano, 2011)

A személyes adatok védelmének (privacy) definícióinak egy része az ember szükségleteit állítja a központba, míg mások a jelen korban is aktuális személyes információ feletti kontrollt. A személyes adatok védelmét az emberi szükségletekre visszavezető meghatározások:

- Az egyének a személyes adatainak védelmére való törekvése egy nem akaratlagos emberi szükséglet, melynek célja, hogy bizonyos tevékenységeinket egyedül végezhessük el, az anonimitás pedig az a fogalom, amikor egy mindenki számára látható cselekedet vagy tevékenység eredményéből harmadik fél számára nem lehet kideríteni az elkövető személyét. (Falkvinge, 2013)
- Xu, Dinev és Smith pedig gazdasági szemszögből írja le a fogalmat. Szerinte a titoktartás alapvető emberi jog: a személy joga az egyedülléthez. Nem abszolút jog abban az értelemben, hogy a mindenkori gazdasági elvek és a költség-haszon közötti kompromisszum. (Xu, Dinev, & Smith, 2011)

Az egyén saját maga személyes adatai felett gyakorolt kontrollt középpontba helyező definíciók:

- France Bélanger és Robert E. Crossler meghatározása szerint az információs magánélet/titoktartás (information privacy) az egyén befolyásolási vagy irányítási képessége a saját adatai felett. (France & Robert, 2011)
- Charles Fries a fogalmat szintén az egyén a saját adatai feletti irányítási képességként definiálja, Alan Westin meghatározása szerint pedig az információs magánélet/titoktartás (information privacy) az egyének, csoportok

és intézmények joga afelett, hogy milyen információt közöljenek mások számára. (Rössler, 2004)

- Roger Clarke szintén hasonló módon definiálja a kérdéses fogalmat: az egyéni érdek, hogy irányíthassa vagy legalább jelentős befolyással bírjon a személyes adatai kezelésére. (Clarke, 1999) Majd az alábbi 4 kategóriába sorolta az információs magánéletet/titoktartást: személyi, viselkedési, kommunikációs és személyes adattal kapcsolatos.

Manapság igen aktuális témának számít az informatikai adatvédelem, ahogy az információ technológia a hétköznapi életünk részévé válik, ezzel együtt a minket körülvevő elektronikus készülékek száma folyamatosan nő, amelyek folyamatosan gyűjtik rólunk az elérhető adatokat. Emiatt – európai értelmezés szerint - az állam egyik feladata, hogy az állampolgárai számára megfelelő arányban biztosítsa az az adatvédelmi jogokat, de ezzel párhuzamosan biztosítsa védelmüket a terrorizmus ellen vagy lássa el a közterületek védelmét vagy a járványok terjedésének megelőzését. Az egyén érdekei egymással ellentétesek lehetnek. (Rössler, 2004)

Ezek miatt Robyn L. Raschke és társai által írt „Understanding the Components of Information Privacy Threats for Location-Based Services" a GPS képes eszközök veszélyeivel foglalkozó cikkében kiemeli, hogy mára a nem informatikai cégeknek is egyértelműen kell kommunikálniuk és jelezniük, hogy miként kezelik és védik az ügyfeleik személyes adatait. Ez az igény annyira nagy, hogy a Szövetségi Hírközlési Bizottság (Federal Communications Commission) kiadott egy legjobb gyakorlatokat összefoglaló riportot a személyes adatok védelmével kapcsolatban. A személyes adatokkal foglalkozó rendszereket már azok tervezési szakaszában fel kell készíteni a lehető legmagasabb fokú védelemre (Robyn, Krishen, & Kachroo, 2014)

Helen Kennedy „Beyond anonymity, or future directions for internet identity research" című cikkében az előző gondolattól jóval tovább megy, a cikkében azt állítja, hogy eljött az idő, hogy az internetes identitás fogalma elmozduljon az „internetes identitás anonim, többszörös és több darabból álló" ténytől az „internetes identitások összeegyeztethetőek az offline énünkkel" tényig. Tehát az online énünk gyakran egybevág az offline-nal, nem pedig bizonyos valós énre jellemző tulajdonságok átformált változatai. (Kennedy, 2006)

Az emberek hozzáállása a magánélet/titoktartás (privacy)-hoz ellentmondásos, mivel szeretnek mások életébe belepillantani, de a sajátjukat már nem szeretik

közzétenni. Érdekes, hogy a közösségi hálózatok megjelenésével a felhasználók tömegének egy újszerű viselkedése jelent meg: akár az életük jelentéktelennek számító vagy akár intim pillanatait is közé lehet tenni. Teszik ezt sokan, a tettük súlyának felmérése nélkül. Ezt hívják **magánélet/titoktartás (privacy) paradoxonnak**. Az egyének féltik a magánéletüket (privacy), de sokszor ezzel ellentétesen cselekszenek. A magánélet/titoktartás (privacy) rizikó az a mérték, amely az egyént éri, ha a személyes adatai kompromittálódnak. (Xu, Dinev, & Smith, 2011)

A további fejezetekben kifejtett kutatásaimban leírom, hogy milyen személyes információk elérhetőek a weboldalak számára az internetes böngészés során. A közösségi hálózatok és kiterjedt hálózattal rendelkező weboldalak sok személyes adatot nyerhetnek ki az egyének online viselkedéséből, amelyeket ők nem feltétlenül explicit módon fedtek fel magukról.

Manapság az egyén személyes adata igen fontos jószággá vált, nincs megbízható kontrollunk afelett, hogy a megosztott vagy közzétett személyes adatok garantáltan eltüntethetőek vagy visszavonhatóak.

2.3 Törvényi szabályozás

Az egyes társadalmak az adatvédelem kérdéséhez különféleképpen viszonyulnak: az Európai Unió direktívái jóval szigorúbbak az Egyesült Államok hasonló tárgyú törvényeinél. Az USA jogszabályozása megengedi az akár már anonimizált számítógépes adatok újbóli személyekhez kötését, az összegyűjtött személyes adatok harmadik fél számára történő átadását vagy az egyénre jellemző böngészőhasználat jellemzőinek gyűjtését. (Rössler, 2004) A különbség abból eredeztethető, hogy a nyugati és az európai társadalmak különféleképpen értelmezik a magánéletet és a szabad akaratot, amelyet legkönnyebben a joggyakorlatokat vizsgálva érthetjük meg:

- Az Amerikai Egyesült Államok törvényei a fő hangsúlyt arra helyezi, hogy az állam az állampolgárokat „békén hagyja”. Itt legfontosabb az egyén szabad akarata.
- Németországban a fő hangsúly azon van, hogy az állam figyelheti-e, amit a polgárai csinálnak vagy mondanak és ha igen, akkor azt milyen szinten. Emiatt a privát szféra védelme akkor merül fel, amikor az emberek a mindennapi életük során már fenyegetve érzik magukat. (Rössler, 2004)

Az uniós joggyakorlat szerint személyes adatok szigorú feltételek mellett gyűjthetőek, kizárólag célhoz kötötten, az adattulajdonos hozzájárulásával.

2016. május 4-én hirdették ki az EU 2018. május 25-én hatályba lépő 2016/679/EU Általános Adatvédelmi Rendeletét (General Data Protection Regulation – GDPR), amely az alábbiakról rendelkezik:

- a személyes adatok törléséhez való jog
- az adathordozáshoz való jog
- az adatvédelmi incidensről való értesülés joga
- a közérthető adatvédelmi magyarázathoz való jog. (Halász, 2016)

A rendelet a 1995-ben hatályba lépett 95/46/EK korszerűsített változata, (Személyes adatok feldolgozása vonatkozásában az egyének védelméről és az ilyen adatok szabad áramlásáról, 1995) és a technológiai fejlődés következtében kialakult új helyzet szabályozására szolgál. (Halász, 2014) Az új szabályozás az Európai Unió tagállamaiban alkalmazott egymástól eltérő adatvédelmi szabályokat igyekszik egységesíteni, ami által teszi lehetővé a szabályozott nemzetközi adatcserét. A szolgáltatókra szigorú előírások vonatkoznak, amelyeket azok kötelesek betartani. A személyes adatokat gyűjtő és kezelő személyek vagy szervezetek kötelesek azokat megvédeni a visszaéléstől. A nem megfelelően eljáró szervezeteket a jövőben keményen szankcionálják. (Európai Bizottság, 2015)

2.4 Az anonimitás keresésének okai

Az online világot használók számára az anonimitás biztosítja, hogy az identitásuk harmadik fél számára ne legyen felfedhető. Az online anonimitásnak különböző szintjei léteznek a névtelen fórumhasználattól a biztonságos fizetésig. Amíg az előbbi esetben a cselekvés eredményét mindenki láthatja, csak a cselekvő személye nem felfedhető, azaz pszeudonimitásról beszélünk, addig a második esetben harmadik fél számára a teljes tranzakció megtörténte titokban kell, hogy maradjon. (Andrews, Gilbert, Repper, Roth, & Wear, 2011)

Az emberiség történelme tele van olyan történetekkel, amelyben szerepet játszott az anonimitás. Rick Falvinge „How Does Privacy Differ From Anonymity, And Why Are Both Important?" című cikkében példaként említi az Amerikai Egyesült Államok kialakulását is. Az egykori angol gyarmat – a jelenlegi Amerikai Egyesült Államok - területén dokumentumokat és röpiratokat tűztek ki a fákra, amelyeken

keresztül buzdították a lakosságot Angliától való elszakadásra. Abban az időben ezért a tettért halálbüntetés járt. Így nem volt kérdés, hogy a tettet végrehajtók miért maradtak anonimak. Így tehát nem létezne az Amerikai Egyesült Államok, ha nem létezett volna anonimitás a Függetlenségi Nyilatkozat aláírása előtt. A magánélet/titoktartás (privacy) és az anonimitás is szükséges a demokratikus közösségben. (Falkvinge, 2013)

Az anonimitás fogalomköréhez szorosan kötődik a **pszeudonimitás** azaz **álnév használata**. Érdekes gondolat, hogy tulajdonképpen a világon senki sem anonim, inkább pszeudonim. (Andreas & Marit, 2010)

Az előbb említett példán kívül is több oka lehet, hogy az emberek miért nem akarják felfedni kilétüket. Ruogu Kang, Stephanie Brown és Sara Kiesler "Why do people seek anonymity on the Internet? Informing Policy and Design" cikkében az anonimitás használata az Internet használatából eredeztethető az alábbiak szerint (Kang, Brown, & Kiesler, 2013):

- ***közérdekű bejelentők:*** a nyilvánosságot nem vállaló, de szabálytalanságot elsőként bejelentők
- ***megbélyegzett csoportok tagjai:*** kisebbségiek, betegek stb.
- ***érzékeny/kényes tartalmakat keresők:*** betegséggel, nemi hovatartozással vagy egyéb érzékeny témával kapcsolatos témájú tartalmakat keresők
- ***hackerek:*** számítástechnikai szakember, aki képes egy számítástechnikai rendszerből lehetetlennek hitt funkciókat előhozni
- ***besúgók/kémek:*** olyan emberek, akik mások számára információkat szolgáltatnak egy személyről

Tehát aszerint, hogy mi a célja az egyes személyeknek az online világban, változik az anonimitás keresésének módja is.

Azonban feltehetjük úgy is a kérdést, hogy miért osztják meg a felhasználók a közösségi hálózaton a személyes adataikat? Mi érdeke fűződik a közösségi hálózatoknak ahhoz, hogy a felhasználóik ezt az ő oldalukon tegyék? Talán a második kérdésre egyszerűbb megadni a választ: Racskó Péter (Szommer, Balogh, & Racskó, 2014) szerint az emberek százmilliói használják az „ingyenes” közösségi hálózatokat, amiért lényegében a személyes adataikkal fizetnek. A közösségi oldalakon található felhasználókat a megadott személyes adataik alapján könnyen lehet profilozni és célzott reklámokat küldeni nekik. A reklámokból befolyt jövedelemből tartják fenn

magukat a közösségi oldalak. A magánélet (privacy) alapvető emberi jog (Smith, Dinev, & Xu, 2011), amiről a felhasználók dönthetnek, hogy milyen mértékben kívánnak élni. Az Internet és a széles körben használt közösségi oldalak miatt felmerülhet a kérdés a magánélet újraértékeléséről és újraértelmezéséről. Az személyes adatok védelme terén az internet fejlődéséhez képest a világ jelentősen le van maradva.

A New York Times – Customer Insight Group kutatócsoportja készített egy felmérést 2011-ben arról, hogy a felhasználók adatmegosztásának okairól. Maga az információ megosztása korántsem újszerű dolog, a 2500 fős kutatás eredményéből kiderül, a közösségi hálózatok megjelenésével az információ megosztására való igény felgyorsult. (Brett, 2011)

A felhasználók információ megosztási kedvének több oka is lehet, az internetet böngésző személy személyiségétől függően változhat az információ megosztásának oka. A kutatás 6 különböző csoportba sorolják az internetezőket az alábbiak szerint:

- Az **önzetlenek** segítőkészek, megbízhatóak és figyelmesek. Többségük e-mail-ben kommunikál és az információt csatolmányként vagy linkként küldik tovább.
- A **karrieristák** intelligens internethasználók, ők azok a közösségi hálózatokban rejlő potenciált elsőként felfedezték. Jellemzően a LinkedIn-t vagy a Facebook-ot használják a kapcsolataik építésére.
- A **hipszterek** fiatalok, népszerűek és szeretik a legújabb technológiát. Ők nem az e-mail-t részesítik előnyben, hanem a kommunikáció gyorsabb és korszerűbb fajtáit, mint az SMS, Twitter vagy a Skype.
- A **bumerángok** azért osztanak meg tartalmakat, hogy visszaigazolást nyerjenek, vagy valamilyen reakciót váltsanak ki. Főként a közösségi hálókat használják, mint a Twitter és a Facebook.
- **A csatlakozók** csoport tagjai kreatívak, figyelmesek és nyugodtak. Ők nagy valószínűséggel e-mail-t vagy Facebook-ot használnak, szeretik az ingyenes termékeket és a promóciókat.
- **A válogatósok** csoport tagjai leleményesek, gondolkodók és óvatosak a megosztott információval kapcsolatosan. E-mail vagy privát üzenet küldését részesítik előnyben a közösségi hálózatokhoz képest. Ezek a felhasználók

tudatában vannak annak, hogy az Internetnek memóriája van, még akkor is, ha azt utólag letöröljük.

2.5 Az anonimitás elhagyásának következményei

Az anonimitás témakörének tárgyalása során említést kell tennünk az anonimitás felfedésének következményeiről. Az internetezéshez használt eszközök, a szoftverkörnyezetük és a látogató online viselkedésének jellemzőinek egy jelentős része hozzáférhető a meglátogatott weboldalak számára. Az összegyűjtött adatból sok hasznos információt lehet kinyerni adatbányászati módszerekkel. Az így kinyert információkból az oldal szolgáltatásait lehetséges fejleszteni vagy a látogatók számára célzott reklámok küldhetőek a webhely által kínált szolgáltatás használata közben.

A kinyert információk segítségével akár teljesen egyénre szabott felületet kaphatnak a személyes preferenciáik által kategorizált látogatók. Azonban a felhasználók elérhető jellemzőik alapján történő kategorizálásának okozata nem feltétlenül pozitív, mint ahogy arra Eli Pariser is felhívta a figyelmet a Beware online „Filter Bubbles" című előadásában, ugyanis nem a vizsgált alany dönti el, hogy a tartalmat milyen megközelítésből szeretné megtekinteni és abba sincs beleszólása, hogy mi az a tartalom, amelyet nem fogyaszthat. Állítása szerint a különböző felhasználói csoportok különböző keresési eredményoldalakat láthatnak a földrajzi elhelyezkedésüknek megfelelően. (Pariser, 2011)

1. ábra: Filter bubble grafikus ábrázolása (Pariser, 2011)

Az információs technológiák fejlődésének következményeként az online világban a valós életünk egyre több aspektusa és a személyiségünk egyre több jellemzője válik hozzáférhetővé. A közösségi hálózatok térhódításával a felhasználók önként osztanak meg magukról tartalmakat, amely elősegítheti a róluk készített profilok pontosabbá tételét. Ezt a jelenséget az angolszász irodalom „privacy

paradox”-nak nevezi: a látogatók elvárják, hogy az adataik védve legyenek az illetéktelenek elől, azonban ezzel ellentétesen viselkednek és felelőtlenül tesznek közzé tartalmakat vagy a személyes adataikat elérhetővé teszik azt kérelmező harmadik felek számára.

3. A LÁTOGATÓK AZONOSÍTÁSÁNAK ÉS KÖVETÉSÉNEK MÓDSZEREI

A látogatók azonosításához és követéséhez a webes böngészés átfogó ismerete szükséges. A fejezet elején leírom a téma megértéséhez szükséges webes böngészés folyamatát. Ezt követően a kutatásaimhoz szükséges saját fejlesztésű adatgyűjtő alkalmazás lefejlesztése előtt összegyűjtöttem a látogató eszközéről, operációs rendszeréről, annak szoftverkörnyezetéről, valamint magáról a látogatóról rendelkezésre álló adatok legszélesebb körét. Ezt követően az előző lépésben összegyűjtött adatok gyűjtésének lehetőségeit vizsgáltam. Majd a *3.4.3 Látogató azonosítása és követése* alfejezetben az összegyűjtött adatokból kiindulva a látogatók azonosításának és követésének lehetséges módszerét írom le.

3.1. A kliensoldal

A weboldalak látogatói a kliensoldalon, a böngészőt használva az internetezésre alkalmas eszközükön tekintik meg a weboldalakat.

3.1.1. Böngészők

A webes tartalmak fogyasztására különböző **hardvereket** használhatnak, melyek tulajdonságaikban jelentősen eltérhetnek egymástól. A világhálón található weblapokat emiatt szokták különféle eszközcsoportokra optimalizálni. A gyakorlatban lehetetlenség lenne minden eszközre külön optimalizálni az weblapokat, ezért megjelenítés előtt az „okos” weblapok és webes alkalmazások lekérdezik a képernyője szélességét, majd annak megfelelő tartalmat jelenítenek meg. Ezeket a szaknyelv responsive design-nak nevezi.

Régebben a különböző böngészők számára a tartalmat a szerveren generálták le, az interneten található weboldalak mobil eszközökre optimalizált verzióját általában az ‘m’ aldomain alatt érhető el, de néhány esetben külön domain alá helyezik át. Manapság a böngésző szélességének függvényében jeleníti meg a böngésző az optimális tartalmat. (Twitter Bootstrap, 2014)

A **böngészők**, mint alkalmazások az internetről letöltött HTML (lásd 8.1 HTML fejezet) oldalakat képesek megjeleníteni. Széles körben elterjedt, weboldalakba ágyazható kliens oldali programozási nyelv a Javascript (lásd 8.3

Javascript fejezet). Az oldalak letöltésének folyamatát részletesen a 8.2 HTTP lekérdezés fejezetben tárgyalom.

Egy felhasználó több eszközt használhat internetezésre. Sőt az sem ritka, hogy egy eszközön több különböző böngészőt használjon valaki. Ez azért fontos tény, mert egy eszközre telepített böngészők nem képesek olvasni egymás kliens-oldalon tárolt adatait (sütiket, localStorage stb.), így ez a felhasználó követését megnehezíti. A látogatók teljes böngészési szokásainak felderítésének fontos része, hogy a felhasználó által használt összes eszközböngésző felderítése és felhasználóhoz való kapcsolása.

A kliens hardverére telepített internetes tartalmak megjelenítésére alkalmas szoftver a **user-agent**, ez leggyakrabban a **böngésző**. HTML nyelvű dokumentumokat képesek értelmezni és megjeleníteni. Több ezer féle böngésző létezik, azonban a felhasználói körük erősen koncentrált. Az internetes társadalom túlnyomó többsége az 5-6 legismertebb böngészőt használja. Az alábbi táblázatban tekinthető meg az 5 legismertebb böngésző elterjedtségte 2017. februárjában.

	Microsoft Internet Explorer/Edge	**Mozilla Firefox**	**Google Chrome**	**Apple Safari**	**Opera**
2017. február	4,8%	15,0%	74,1%	3,6%	1,0%

2. táblázat: 5 legelterjedtebb böngésző **(W3Schools, 2017)**

A különböző böngészők különböző megjelenítő motorokkal rendelkeznek, amelyek eltérő módon jelenítik meg a weblapokat és különbözőképpen hajtják végre a Javascript utasításokat.

Böngésző	Megjelenítő motor (layout engine) neve
Microsoft Edge	EdgeHTML
Microsoft Internet Explorer	Trident
Mozilla Firefox	Gecko
Google Chrome	Blink
Apple Safari	WebKit
Opera	Presto

3. táblázat: Néhány ismertebb böngésző és megjelenítő motorja **(Stanclift, 2008)**

Az inkognitó mód

A böngészők többségében a 2008-2010 között jelent meg az **inkognitó mód**[12]. Normál használat közben a böngészők többek között az alábbi adatokat mentik merevlemezre:

- meglátogatott weblapok tartalma (cache)
- meglátogatott weblapok időpontja
- cookie-k, localStorage és sessionStorage tartalma

A felhasználó által meglátogatott weblapok tartalmát a böngésző a merevlemezre menti, mert a weboldal későbbi meglátogatása esetén csak a módosult elemeket kell letölteni, ezáltal tehermentesíthető a hálózat, valamint a weboldal megjelenítése is sokkal gyorsabb lesz. A böngésző normál üzemmódját használva a böngészés befejeztével a merevlemezre lementett adatokból visszanyerhető a felhasználó böngészési előzménye. Ezt elkerülhető az inkognitó mód használatával, mert ebben az esetben a böngésző a munkamenet befejeztével nem hagy nyomot a merevlemezen, tehát olyan, mintha a böngészés meg sem történt volna. A böngésző bezárásával a sütik törlődnek a localStorage pedig a sessionStorage-hoz hasonló viselkedést mutat és az is szintén törlődik a munkamenet lejártával. Fontos megjegyezni, hogy az inkognitó mód használatával a weboldalak ugyanúgy követhetnek minket, számukra minden adat és azonosítási faktor ugyanúgy elérhető, a bejelentkezés nélküli böngészés esetében is.

HTTP fejléc

[12] böngészőnként máshogy hívják: Internet Explorer-ben InPrivate mode, Mozilla Firefox-ban Private browsing, Google Chrome-ban incognito mode stb.

A böngészők a HTTP lekérdezések fejlécében az 2. ábra látható elemeket küldik el, melyek közül az egyes elemek jelentése rendre:

- **Accept:** a böngésző által kezelni képes állományok és prioritásuk
- **Accept-encoding:** a böngésző által támogatott átvitel során használható tömörítési algoritmusok
- **Accept-Language:** a böngésző által támogatott nyelvek
- **DNT (Do-no-track):** ha a mező értéke 1, akkor a meglátogatott weblap nem fogja követni a felhasználót
- **Host:** a lekérdezni kívánt URL
- **User-Agent:** a weboldal megjelenítéséhez használt eszköz, a böngésző és operációs rendszer típusának és verziójának megállapítására szolgáló szöveg, jelen esetben az operációs rendszer típusa Windows 7, 64 bites verzió, a böngésző pedig Mozilla Firefox 22 (Gecko), a hardverről nem tartalmaz semmi extra adatot

Accept	text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8
Accept-Encoding	gzip, deflate
Accept-Language	en-US,en;q=0.5
Connection	keep-alive
DNT	1
Host	index.hu
User-Agent	Mozilla/5.0 (Windows NT 6.1; WOW64; rv:22.0) Gecko/20100101 Firefox/22.0

2. ábra, az index.hu HTTP lekérdezés fejléce (saját felvétel)

A User-Agent string felépítése a következő:

```
Mozilla/x.0 (operációs rendszer jellemzői) megjelenítő motor
        (böngésző jellemzői) telepített szolgáltatások
```

Az alábbi jellemzőket foglalja magában:

- Mozilla megjelenítő motorral való kompatibilitás
- A böngészőt futtató operációs rendszer jellemzői
- A böngésző megjelenítő motorjának neve és verziója
- Böngésző neve és verziója
- Telepített szolgáltatások és egyéb jellemzők (Andersen, 2008) (Shall, 2017)

DNT

A 2. ábrán is látható, hogy a böngésző elküldött egy DNT paramétert a HTTP fejlécben. Ez a **Do-Not-Track** kifejezés rövidítése, ami magyarul annyit jelent: "ne

kövess" (Mozilla.org). Bevezetését a W3C[13] 2009-ben indítványozta. Ha látogató úgy kívánja, hogy az általa meglátogatott weblap vagy egyéb harmadik fél ne kövesse őt, a böngésző a HTTP lekérés fejlécében elküldi ezt az információ a szervernek. Mára az ismertebb böngészők támogatják a DNT fejléc küldését, azonban a felhasználók nagy része nem tud róla, mivel nem reklámozzák ennek lehetőségét a böngészők. 2012 decemberében a Firefox felhasználók 90%-a nem használta még ezt a funkciót. (Mathews, 2013) A kapcsoló hatását a felhasználó nem tudja ellenőrizni, hiszen nem kap semmiféle visszajelzést, hogy az általa meglátogatott weblap bármit is lementett róla. A Yahoo! és a Twitter támogatja ezt a funkciót. Azonban a felhasználó semmilyen visszajelzést nem kap arra vonatkozóan, hogy a weblap támogatja-e ezt a funkciót, valamint, hogy azt valóban be is tartja-e.

A DNT jövője megkérdőjelezhető, mivel a valódi célját nem éri el a felhasználó a használatával, hanem pont az ellenkezőjét, ugyanis a HTTP fejlécben átküldött DNT is hozzájárulhat egy eszközböngésző beazonosításához. A Yahoo! elsőként kezdte el támogatni a DNT-t, azonban 2014. májusában a Yahoo! felfüggesztette azt. (The Stanford Review, 2014)

3.1.2. A munkamenet

A HTTP szerver nem jegyzi meg a kliens adatait az egyes lekérdezések kiszolgálása között. Ezt a rést hivatott kitölteni a **munkamenet**. A kliens minden HTTP lekérdezés fejlécében elküldi az adott domain-hez tartozó érvényes munkamenet sütiket. Első látogatás alkalmával, amikor a szerveroldalon indítjuk a munkamenetet a szerver generál egy 27-40 karakter hosszúságú betűkből és számokból álló munkamenet azonosítót, amelyet visszaküld a kliens számára. Ez lesz a munkamenet azonosítója, amelyet a PHP alapesetben PHPSESSID nevű sütiben tárol. A süti egy kliens oldali adattároló technológia, bővebb leírása a *Süti (cookie)* fejezetben olvasható.

A szerveroldalon egy kitüntetett könyvtárban létrehoz egy file-t az előzőleg kiosztott munkamenet azonosító néven, a tartalma pedig a munkamenetbe elmentett változók és azok tartalma lesz. Lehetőség van akár tömbök és objektumok tárolására is, ebben az esetben szerializálva[14] lesznek elmentve. A következő oldalletöltés

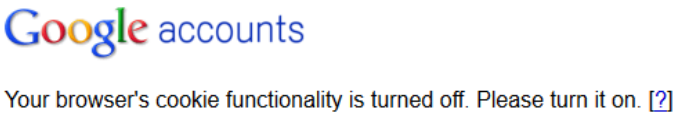
[13] World Wide Web Consortium: nemzetközi internetes szabványügyi konzorcium
[14] Memóriabeli objektumok szöveges formában történő ábrázolása

alkalmával a munkamenet azonosítónak megfelelő munkamenet file tartalma rendelkezésre fog állni a programozási környezetben.

A kliensoldalon a munkamenet azonosító egy sütiben tárolódik. Amikor a látogató ismételten meglátogatja az oldalt, amelyhez a süti tartozik, a böngésző a lekérdezés fejlécében elküldi azt a szerver számára.

Ha egy támadó megszerzi egy bejelentkezett felhasználó munkamenet azonosítóját és bemásolja a saját gépén a weboldal által készített süti file-ba, akkor hozzáférhet az áldozata által látott tartalmakhoz. Ez természetesen csak abban az esetben működik, ha a weboldalt nem készítették fel a **session fixation** támadás ellen. A védekezés egyszerű, csak ellenőrizni kell minden lekérdezésnél a látogató IP címét és User-agent stringjét. Ha valamelyik eltér az előző lekérdezésben meg, a felhasználót ki kell léptetni.

Mi történik olyan esetben, ha a kliens letiltja a sütiket? Ha a weblap megírója nem végzett alapos munkát, akkor a felhasználók nem fognak tudni bejelentkezni, mivel a lekérdezések fejlécében nem lesz benne a munkamenet azonosítója.



3. ábra, A GMail nem működik sütik használata nélkül (saját szerkesztés)

Az olyan weboldalak, amelyek esetében követelmény, hogy sütiket nem támogató böngészőkben is futtathatóak legyenek, (pl: banki rendszerek) a query string használható a munkamenet azonosító átküldésére.

3.1.3. A kliens oldali adattárolás

A HTML5 egyik nagy újítása a kliens-oldali adattárolási lehetőségek kibővítése. Több fontos különbség is van - az elődje – a sütik és az Web Storage technológiák között, hogy az utóbbiak mérete a sütik méretének sokszorosa (5-25 MB a böngésző fajtától függően) is lehet, valamint a tárolt adatok a lekérdezések során nem lesznek elküldve a szervernek.

A **localStorage** és **sessionStorage** a sütikhez hasonlóan kulcs-érték párok tárolására szolgáló technológia kliensoldali technológia. Amíg sessionStorage csak a munkamenet idejére jegyzi meg a beleírt adatokat, a localStorage-ban tárolt adatok a

böngésző bezárása után is használhatóak maradnak. A localStorage előnye, hogy egyelőre kevés olyan alkalmazás ismeri, amellyel a felhasználók a böngészési gyorsítótárat törölheti le, így az abban tárolt adatok hosszabb időn át maradnak elérhetőek. Mára széles körben elterjedt technológia.[15]

Egy másik adattárolásra alkalmas technológia a **WebSQL**, ami a böngészőbe implementált SQL adatbáziskezelő. Jelenleg a Chrome, Opera, Safari és kiegészítővel a Firefox is támogatja, az IE nem.[16]

Az **application cache** szintén a HTML5 egyik újdonsága, amely arra hivatott, hogy a weblapok offline állapotban is nyújthassanak funkcionalitást. Amíg a webes alkalmazás online állapotban van, letölti az offline működéshez szükséges részeket. [17]

A **SharedObject**, Flash sütinek is hívott kliens oldalon tárolt adatcsomag. Működése nagyon hasonló a normál sütiéhez, azonban ezt csak Flash alkalmazásokból lehet elérni, viszont mérete nem csak 4 kB lehet, hanem alapból 100 kB. Ha a kliens gépén futó alkalmazásnak ennél is több helyre van szüksége, akkor az alkalmazás először kérvényezi a tárterület megnövelését a felhasználótól, aki saját belátása szerint engedélyezheti vagy elutasíthatja azt. A SharedObject nagy előnye, hogy azt egyszerre több Flash alkalmazás is tudja használni és annak változása esetén értesítést kapnak róla.

Etag vagy **Entity Tag** a HTTP protokoll része. Tulajdonképpen egy mechanizmus, amely segítségével a böngésző megállapíthatja, hogy az ismételten letölteni kívánt tartalom frissült-e a szerveren. Ha a böngésző gyorsítótárában található file nem frissült a szerveren, akkor felesleges azt újra letölteni, így ezzel értékes idő és sávszélesség takarítható meg. Egy file gyorsítótárazása során a böngésző elmenti a file módosításának dátumát. Így annak a kiderítése, hogy a file frissült-e a szerveren az, ha összehasonlítjuk a szerveren lévő file módosításának dátumát a letöltött file módosításának dátumával. Ha a szerveren lévő korábbi dátum, akkor le kell tölteni, mert a tartalma módosult. Az ETag-et a különféle webszerver alkalmazások különféleképpen számíthatják. Általában ez egy hash, amelyet a letölteni kívánt file tartalmából számítanak. Így, ha a letölteni kívánt tartalom utolsó módosításának dátuma nem frissülne, ez a hash akkor is változik.

[15] http://caniuse.com/#search=localstorage
[16] http://caniuse.com/#search=websql
[17] http://caniuse.com/#search=offline

Az előbbiekben ismertetett kliens oldali adattároló technikák használatával a weboldalak képesek egy adott eszközt azonosítani, így alkalmasak a látogatók követésére. A több oldalba beillesztett komponensek révén az adatgyűjtő weboldal képes egy adott eszköz által meglátogatott weboldalak listájának előállítására.

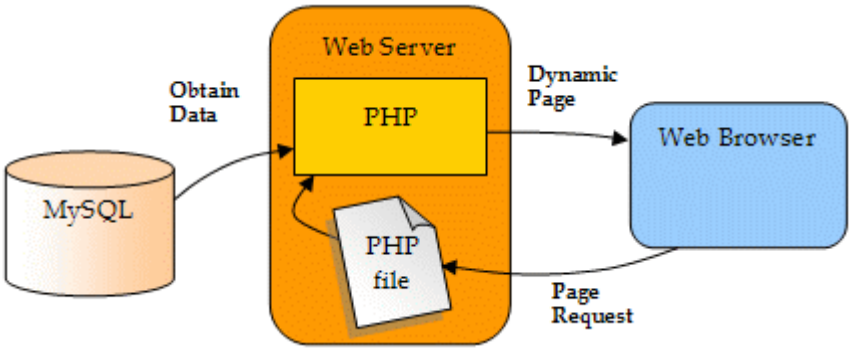
A szakirodalom **browser fingerprint** néven szokott hivatkozni a böngészőből kiolvasott bizonyos paraméterekből képzett hash-re. Az Electronic Frontier Foundation (http://panopticlick.eff.org) oldalán az alábbi adatokból számítják a hash-t, ami a felhasználó által használt eszközböngészőből lekérdezhető jellemzőkből alkotott adatcsoport, amely rövid időtávon eszközböngésző és a használójának azonosítására használható:

- user-agent string (HTTP_USER_AGENT)
- böngésző által kezelni képes állományok és prioritásuk (HTTP_ACCEPT)
- böngészőre telepített kiegészítők és verziójuk
- időzóna
- képernyő mérete és színmélység
- telepített betűkészletek
- sütik engedélyezve vannak-e?
- localStorage/sessionStorage és IE.userdata teszt (IE userdata a Microsoft által fejlesztett a localStorage-hez hasonló kliensoldali adattárolási technológia, IE5-től IE8-ig támogatott)

4. ábra, követésre alkalmas technológiák fejlődése (saját)

3.2. A szerveroldal

Az 5. ábra látható módon a böngésző a webszerverrel kommunikál, az pedig a lekért erőforráshoz rendelt programozási nyelvet hívja meg, hogy állítsa elő a kívánt tartalmat. A programozási nyelv - jelen esetben PHP - szükség esetén az adatbázishoz fordul az adatokért, majd a programkódban leírt módon feldolgozva azt előállítja a HTML kimenetet, amelyet a webszerver visszaküld a kliens számára.



5. ábra: webszerver kapcsolata az adatbázissal és a böngészővel (Bhavin, 2014)

Az internetes kommunikációs szabályokat a **HTTP - Hypertext Transfer Protocol** – foglalja magában, ami az ISO/OSI és a TCP/IP alkalmazási rétegének webes kommunikáció átvitelére alkalmas alprotokollja. A protokoll kérdés-válasz alapon működik, ami azt jelenti, hogy általában a kliens küld egy kérést a szervernek, ami előállítja számára a választ és visszaküldi. A HTTP egy állapot nélküli protokoll, így az egy felhasználótól érkező lekérdezésekről a szerver alapesetben nem tárol információt vagy állapotot. Szükség esetén ez mégis lehetséges sütikkel, rejtett űrlap változók használatával vagy a query string-en keresztül.

```
GET / HTTP/1.1
Host: index.hu
Connection: keep-alive
Cache-Control: no-cache
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8
Pragma: no-cache
User-Agent: Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/28.0.1500.95 Saf
ari/537.36
Accept-Encoding: gzip,deflate,sdch
Accept-Language: hu-HU,hu;q=0.8,en-US;q=0.6,en;q=0.4
Cookie: PHPSESSID=a42e6eaafb621c7a79208a4cc4f36514; ident=520dc38c8a86173626000730; cikkvegi_ajanlo1=cimlapi_a
janlo; WACID=1376753932000A65131536; index_mobile=false; INX_CHECKER2=1
```

6. ábra, HTTP lekérdezés fejléce (saját felvétel)

A 6. ábra egy HTTP lekérdezés fejléce látható. Első sorában a lekérdezés metódusa látható, jelen esetben a GET metódust használja és a HTTP 1.1-es protokoll szerint zajlik a kommunikáció. A következő sorban látható a host, ahol elérhető a letöltendő adat. Majd több más paraméter közt az Accept a támogatott formátumokat, a User-Agent-et, a támogatott nyelveket és a sütiket jelöli.

A TOR (The Onion Router) egy anonimitást biztosító rendszer, amely böngésző és a webszerver közti kapcsolatot titkosítva biztosítja a küldő anonimitását.

A rendszer a több szinten különböző titkosítással látja el a küldő csomagjait, megakadályozva a küldő pozíciójának kiderítését. (The Onion Router, 2015) Helyes használatával a weboldalak nem képesek a visszakövetni a látogatóikat, kivéve, ha megadják személyes adataikat.

3.3. A böngészés folyamata

A jelen dokumentumban **felhasználó azonosításnak** nevezem azt a folyamatot, amikor a felhasználó által használt eszközböngészők és a felhasználó viselkedéséből kikövetkeztetve ismertté válik valamilyen egyedi azonosítója (a felhasználó személye, rendszer azonosító (Facebook ID), e-mail cím stb.) vagy az üzlet szempontjából fontos tulajdonságai alapján profilozhatóvá válik.

Felhasználó követésnek nevezem azt a folyamatot, amikor két weboldal letöltésről egyértelműen megállapítható, hogy ugyanaz a felhasználó indította a lekérdezést. Ebbe beletartozik az eszközböngészők egymáshoz kapcsolása is.

Az egy weblapon történő böngészésnek két fontos momentuma van:

- a weblap első megtekintése első alkalommal
- a weblap megtekintése a második vagy sokadik alkalommal

3.3.1. A weblap megtekintése

Egy weblap első és sokadik alkalommal történő megtekintése között az a különbség, hogy első alkalommal még nincs elindítva a munkamenet, így a lekérdezés fejlécében nem lesznek a munkamenet azonosítók elküldve a szervernek, így az egy felhasználótól független kimenetet fog visszaadni. Az első válaszüzenetben a szerver küldi vissza a kliens számára a munkamenet azonosítót, amelyet a böngésző elment.

A második vagy a sokadik alkalommal történő megtekintés során a lekérés fejlécében a böngésző elküldi a szerver számára az oldalhoz tartozó még nem lejárt sütiket, ami a szerveroldalon a kérés megérkezését követően a meghívott programozási nyelvben rendelkezésre fog állni, amennyiben munkamenet azonosító volt, akkor a munkamenet változók állnak rendelkezésre.

3.3.2. A böngészés révén kinyerhető adatok

A látogatók beazonosításának és követésének legkézenfekvőbb módja, ha ezeket az adatokat maga a látogató bocsájtja az azt igénylő weblap számára. Ha elvetjük azt az esetet, amikor a látogató a regisztrációt követően mások számára is

elérhetővé teszi a belépéshez szükséges azonosítóját és kódját, akkor a weblap minden esetben követni tudja őt, valamint elérhető lesz számára a látogató néhány adatát: időpont, IP cím, IP címből visszafejtett földrajzi pozíció, a webhely megtekintett weblapjai, HTTP fejlécből kinyerhető adatok és a DOM-ból kinyerhető adatok.

A regisztráció folyamán megadott adatok valódiságát természetesen nem lehetséges ellenőrizni. Ha nem ellenőrzi a webhely, akkor a látogatók hajlamosak hamis e-mail címek megadására. Emiatt szoktak sok esetben egy visszaigazoló e-mail-t kiküldeni a megadott e-mail címre.

A felhasználó eszköze és a weblap szervere között történő kommunikációban az alábbi aktorok lehetnek érintettek:

- **munkáltató/ISP (Internetszolgáltató):** segítségével kapcsolódunk a világhálóra. Ez a szervezet adatokat szerezhet a hozzá csatlakozó felhasználókról.
- **weblapok:** a böngészés során megtekintett weblap információkat szerezhet a felhasználóról
- *hirdetők és közösségi hálózatok:* a weboldalakba helyezett nyomkövető kód révén információkhoz juthat a weblap látogatóiról.
- **hackerek és kiberbűnözők:** különféle módszerekkel adatokat szerezhetnek a weblapok látogatóiról
- **állam:** Az állam és az állami hatáskörbe tartozó szervezetek adatszolgáltatásra kötelezhetik a szolgáltatókat
- **köztes csomópontok:** a kommunikáció kezdő és végpontja közötti bármely csomópont, amelyen keresztülhaladnak az adott kommunikációhoz tartozó csomagok

A listából kitűnik, hogy bizonyos szereplők többletinformációkhoz is hozzáférhetnek az adott látogatóról. A kutatásom során csak a bárki által hozzáférhető adatokra fókuszáltam, tehát a felsorolt aktorok közül többnyire a weblapok, kisebb részt a hirdetők és a közösségi hálózatok a kutatásom alanyai. Ezek az aktorok ugyanis csak a böngésző, az operációs rendszer és a protokoll által lekorlátozott adatokat érhetik el. Nincs hozzáférésük például olyan adatokhoz, hogy melyik ügyfél azonosítóhoz kiosztott IP címről milyen lekérdezések érkeztek. Ezzel a rendelkezésre álló adatok csak egy részhalmaza érhető el, valamint az azonosítás nem minden esetben lesz triviális, de ezek az adatok szinte bárkiről elérhetőek.

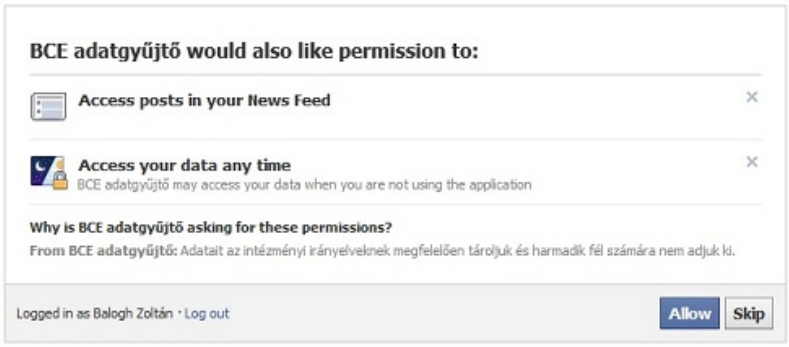
Felhasználó azonosítása regisztráció esetén

Bár a felhasználó bejelentkeztetésének több módja is lehet, minden esetben a lényeg, hogy a látogatót egy munkamenet azonosítóval lássuk el, aminek következtében a szerveren be tudjuk azonosítani. Az alábbiakban bemutatom, hogy milyen módjai vannak a látogató bejelentkeztetésének:

- **Űrlapon keresztül:** A bejelentkezésre szolgáló weblapra egy bejelentkező űrlapot helyeznek el, az űrlap tartalma a bejelentkezés gombra történő kattintást követően általában POST típusú – az átlagfelhasználó számára nem látható módon - lekérdezésben küldi el a böngésző a szervernek.
- **Közösségi hálózatok vagy OpenID révén:** Harmadik félnél tárolt személyes adatok átadása néhány kattintással az azt igénylő webhely számára

A felhasználók azonosítása a közösségi hálózatok által

Facebook Connect, Google+ Sign-In, Sign in with LinkedIn stb. a közösségi hálózatok egyik népszerű szolgáltatása. Segítségével a megfelelő bejelentkező komponenst tartalmazó weboldalon néhány kattintással regisztrálni lehet a weboldalon. Nem szükséges külön regisztrálni és a személyes adatainkat kézzel megadni, elég pl. az oldalon lévő Facebook-os regisztráció, majd a Facebook-os belépés gombra kattintani. Ezt követően egy felbukkanó ablakban megjelenik a jól ismert Facebook bejelentkezési oldal, ha a felhasználó éppen nincs bejelentkezve, majd rákérdez, hogy a weboldal milyen adatokhoz férhet hozzá. Többek között hozzáférést lehet igényelni a felhasználó alapadataihoz, úgymint a neve, születési helye, tartózkodási helye, neme, profilképe stb. De lehetőség nyílik hozzáférni akár a felhasználó képeihez, barátlistájához vagy akár falához is (Facebook Developers, 2013). 2012. őszétől kezdve az weblapok és alkalmazások több különböző szintű hozzáférést is kérhetnek a felhasználók adataihoz. Ez azt jelenti, hogy a csatlakozni kívánó egyén számára nem feltétlenül szükséges az alkalmazás által kért össze hozzáférést megadnia. A felhasználónak lehetősége van kiválasztani, hogy mihez engednek hozzáférést.



7. ábra, Egy harmadik fél által írt alkalmazás engedélyeket kér a Facebook-tól (saját felvétel)

Az adatok titkosításához és a hitelesítéshez a Facebook az OAuth 2.0 nyílt szabványt használja, melynek révén nyílik lehetőség a harmadik fél által fejlesztett alkalmazások számára a hozzáféréshez, anélkül, hogy a webalkalmazásnak bármilyen saját kulcsot vagy kódot ki kelljen adnia.

Az OpenID egy nyílt szabvány, amely a fentebb leírt közösségi oldalas regisztráció és bejelentkezésre hasonlít. A kettő közötti fő különbség abban rejlik, hogy ez az informatikai iparág nagyjai által létrehozott független bejelentetési szabvány keretrendszer. Hátránya, hogy kevésbé elterjedt, hozzávetőlegesen 190,000 weboldalon használható. (Built With, 2013)

A weboldalak az űrlap és a Facebook Connect vagy Google+ Sign-in kombinációját szokták alkalmazni, néhány esetben lehetőséget biztosítanak az OpenID-vel rendelkező felhasználók számára is. Jelenleg egyik módszer elhagyása sem javasolt, mivel, ha az oldalon csak Facebook Connect regisztráció található, akkor a Facebook azonosítóval rendelkező nem rendelkező felhasználók nem fogják használni az oldalt.

3.3.3. A rendelkezésre álló adatok

Elméleti szinten rengeteg hozzáférhető adat létezik egy weblap számára, amelyek az alábbi kategóriákba egyikébe sorolhatóak:

- a böngészéshez használt **hardver** tulajdonságai
- a böngészéshez használt hardveren lévő **szoftver környezetének tulajdonságai**
- a böngészéshez használt **böngésző** tulajdonságai
- a felhasználó **viselkedése** vagy a **felhasználóra jellemző tulajdonságok**

- **harmadik fél** által történő azonosítás

Az Internet böngészéséhez használt - a hardver és az operációs rendszer felett elhelyezkedő - célalkalmazás a böngésző. A böngésző a webszervertől letöltött HTML állományok letöltésére és megjelenítésére szolgál.

8. ábra, a látogató és az Internet közötti kapcsolat (saját szerkesztés)

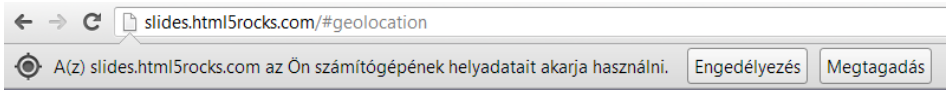
A 8. ábra, *a látogató és az Internet közötti kapcsolat* látható módon éri el a felhasználó az Interneten található weboldalt, a köztes szoftver és hardver elemeket használja. Ezek az elemek egymás számára szolgáltatásokat nyújtanak és vesznek igénybe. A weblapok számára minden egyes elem attribútumainak egy része elérhető, némely tulajdonságokat pedig szándékosan elrejt a következő réteg.

3.3.4. A hozzáférhető adatok

Az eddig bemutatott adatok elviekben hozzáférhetőek, azonban a gyakorlat azt mutatja, hogy bizonyos adatokhoz a böngészők sem férnek hozzá vagy azt szándékosan nem teszik hozzáférhetővé a weblapok számára. Ez természetesen böngészőnként és platformonként is változhat. (pl: az Apple termékek nem támogatják a Flash lejátszókat, így annak verziószáma, valamint a rendszerre telepített betűtípusok listája nem lesz elérhető vagy bizonyos HTML5 képességek a böngészők[18] asztali verziójában szintén nem elérhetőek).

A személyes jellegű adatok esetében felhasználói beleegyezés szükséges azok eléréséhez. (pl: GeoLocation, közösségi hálózatról hozzáférhető adatok) Általában ezeket a látogatók nem szívesen osztják meg a weboldallal.

[18] nem elérhető a böngészéshez használt kapcsolat típusa



9. ábra, a Chrome engedélyt kér a látogató földrajzi pozíciójához (saját felvétel)

Az inkognitó mód (3.1.1/Az inkognitó mód) detektálása régebben a „CSS history sniffing" technika segítségével volt lehetséges (Stack Exchange, 2011), azonban ezt a biztonsági rést mára már az összes böngészőben javították, így ennek detektálására már nincs lehetőség. Korábban a következő módon volt lehetséges annak detektálása: normál esetben, ha a látogató egy linkre kattintva, majd a vissza gombra kattintva ugyanazon az oldalon lévő link színe megváltozik, jelezve a felhasználó számára, hogy a link mögött lévő oldalt a felhasználó már egyszer meglátogatta. Ha ezt a műveletet Javascript segítségével idézte elő az alkalmazás fejlesztője és a végén lekérdezte, hogy az adott link színe megváltozott, akkor máris megkapja a választ arra a kérdésre, hogy a látogató inkognitó módban használja-e a böngészőjét. A dolog működésének megértéséhez tudni kell, hogy az inkognitó mód célja az, hogy miután a felhasználó bezárta a böngészőt az eszközön ne maradjon nyoma a böngészésének. Azonban ezt a hibát mára már befoltozták, így nincs lehetőség az annak detektálására, azaz nem lehet a felhasználót arra kényszeríteni, hogy ne használjon inkognitó módot és nyoma maradjon a gépen és ebből kifolyólag követni lehessen a későbbiekben. Érdekességként jegyzem meg, hogy bár az inkognitó mód használatával a böngésző minden nyomot megpróbál eltüntetni a gépről, a meglátogatott domain-ek listája (Windows alapesetben 5 percig) a DNS resolver gyorsítótárában megtalálható az eszközön.

A függelék 8.6 fejezetében található táblázatban foglaltam össze a lényegesebb paraméterek használatának előnyeit és hátrányait.

A HTML5 és CSS3 képességek jelentős részét alap Javascript segítségével is lehet detektálni, azonban a Modernizr JS library segítségével egységesített formában egy objektum attribútumainak lekérdezésével érhető el. A különböző beépülő modulok detektálása és verziójának lekérdezését szintén célszerű egy erre szakosodott library-val végeztetni. Mivel az egyes képességek teljes mértékben összefüggenek a különböző böngészők és azok verzióival, így azok nem járulnak hozzá a felhasználó pontosabb azonosításához, emiatt ezen adatok lementése csak akkor szükséges, ha az oldal látogatóinak böngészőjének képességeiről szeretnénk statisztikát készíteni.

A billentyűzet és egérmozgás figyelése a kliensoldalon erőforrás-igényes művelet, ezért a gyakorlattban az összegyűjtött adatok tömörítésére lehet szükség a szerverre való átküldés előtt.

A harmadik féltől lekérdezhető adatok jellemzőit a *4. táblázat: harmadik féltől lekérdezhető paraméterek előnyei és hátrányaban* foglaltam össze. Az adatok közös jellemzője, hogy API-n keresztül a felhasználó vagy az általa használt eszköz egy tulajdonságával kérdezhetőek le.

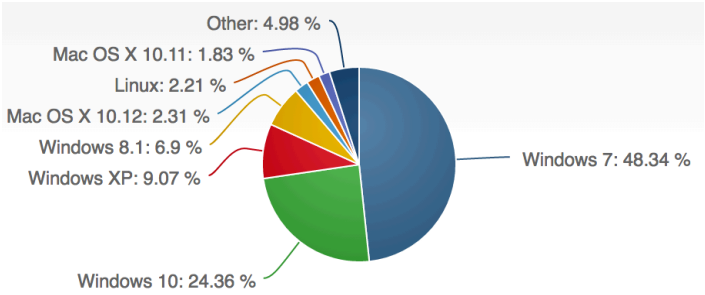
Megnevezés	Előny	Hátrány
ISP	kikövetkeztethető belőle az internetezéshez használt hálózati kapcsolata, bizonyos esetekben a kapcsolathoz az internetet szolgáltató intézmény	
Proxy	ha az illető proxy-t használ a kapcsolathoz a használt IP címe csak a munkamenet azonosítására használható, mert az nem a valós	NAT mögötti eszköz detektálására nem használható
Blacklist check / tiltólistán való jelenlét	megtudható, hogy az adott IP cím valamelyik spam tiltólistán szerepel	
Közösségi oldalakon hozzáférhető adat	rengeteg hasznos felhasználóra jellemző információt lehet letölteni	felhasználói hozzájárulás szükséges a hozzáféréshez, nehéz rábírni a felhasználókat adataik megadásához
Egyéb harmadik féltől lekérdezhető adat	szolgáltatótól függően az adatok személyhez vagy eszközhöz köthetőek	ha a lekérdezés kulcsa az IP cím, nem biztos, hogy a lekérdezés releváns találatot szolgáltat

4. táblázat: harmadik féltől lekérdezhető paraméterek előnyei és hátránya

A lekérdezendő adatok megállapításához célszerű tesztelni, hogy mely paraméterek milyen mértékben járulnak hozzá az egyedi böngésző ujjlenyomat előállításához. Abban az esetben, ha valamelyik paraméter egy másikkal korrelál (mint a böngésző típusa, verziója és a HTML5, CSS3 képességei) az elhagyható. A

gyakorlatban az azonosítást és felhasználó követését lehetővé tevő lekérdezhető adatok körének megállapítása üzleti érdekek és technológiai megvalósíthatóság mentén létrejövő konszenzus.

- A **környezeti változóként vagy Javascript segítségével elérhető adatok** csoportjába tartozó elemek (aktuális böngészőablak méretének kivételével) mindegyikét célszerű lekérdezni, mivel elérésük nem hardverigényes és az eszközböngésző beazonosításában kulcsszerepet játszanak, valamint ezen adatok segítségével érhetőek el a harmadik félnél tárolt adatok többsége. A generált eszközböngésző azonosító teszi teljes mértékben bizonyossá, hogy az aktuális eszközböngészőről már meglátogatták a weboldalt.
- A **felhasználóra jellemző** billentyűzet és egérfigyelés előnye, hogy felhasználót képesek azonosítani, azonban a fejlesztés során arra is gondot kell fordítani, hogy az egér és a billentyűzethasználatból származó adatok az oldalletöltések között is megmaradjanak. Szintén az azonosítás pontossága érdekében az adatgyűjtés többször érdemes elvégezni, valamint biztosítani kell, hogy a felhasználó megfelelő mennyiségű mintát adjon le. Kerülni kell azokat az eseteket, amikor a felhasználó valamilyen eseményre való reakciójaként nyom le egy billentyűt vagy kattint az egérrel, (pl: játék közben) mert ezek nagyban ronthatják az azonosítás pontosságát.
- A **komplex paraméterek** (telepített betűkészletek, billentyűzet és egérhasználat) csoportjába tartozó elemek erőforrás- és sávszélesség-igényesek lehetnek, ezért nem minden esetben javasolt a használatuk. A telepített betűkészletek listájának lekérdezéséhez Flash lejátszó szükséges, amely az Adobe állítása szerint a desktop PC-k 99%-án található Flash lejátszó. (Adobe, 2011) Más források 92-95% közé teszik az elterjedtségét földrajzi hely függvényében. Az operációs rendszerek piaci részesedése a *10. ábra* látható, amiből az derül ki, hogy a Mac OS X elterjedése 6,65%-os, tehát ezeken a gépeken biztosan nem lehetséges a betűkészlet detektálása, mivel ezek nem támogatják a Flash lejátszót.



10. ábra, asztali operációs rendszerek piaci részesedése 2016-ban (NetMarketShare, 2016)

- A **harmadik féltől lekérdezhető adatok** egy részének (proxy, blacklist check és egyéb harmadik féltől lekérdezhető adat) pontossága megkérdőjelezhető, tehát semmiképpen nem kezelhető biztos adatforrásnak abban az esetben, ha annak hitelességét más forrás nem erősíti meg. A közösségi hálózatokon található adatok megosztásához engedélyt kell kérni először a látogatótól, személyes tapasztalataim azt mutatják, hogy a felhasználók csak csekély százaléka engedi egy külső oldal számára, hogy a Facebook-os adataihoz hozzáférjen. A Brafton felmérése szerint azonban a jelenlegi tizenéves generáció 91%-a rak ki képeket a Facebook-ra rendszeresen, ugyanez a mutató 2006-ban 69% volt. (Brafton Editorial, 2013) Ami valószínűsíthetően magasabb hozzájárulási hajlandósággal jár együtt.

Az adatokat többféle forrásból nyerhetjük ki:

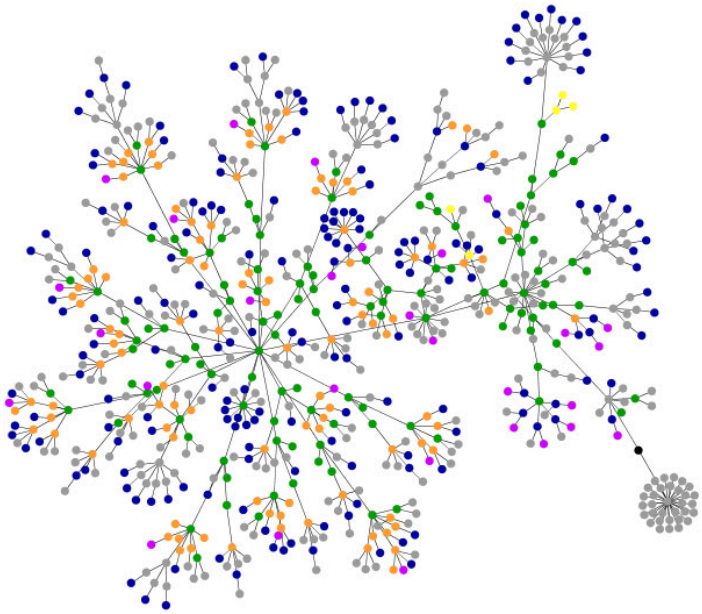
- **A HTTP fejlécben elküldött adatok:** minden egyes lekérdezéssel együtt a lekérdezés fejlécében a szerverre utazik az eszköz hardverének és operációs rendszerének bizonyos paraméterei
- **A Javascript segítségével kiolvasható tulajdonságok:** a böngésző DOM-jából Javascript segítségével olvasható ki, majd ezt követőn AJAX lekérdezéssel lehetséges a szerverre elküldeni azt
- **A harmadik féltől megszerezhető adatok:** valamilyen eszközre vagy felhasználóra jellemző tulajdonság egy harmadik fél által történő lekérdezése a szerveren

3.3.5. A felhasználói életút vizsgálata

Az ugyanazon látogatók egy munkamenetéhez tartozó lekérdezéseinek összekapcsolása egyszerű, hiszen minden lekérdezés fejlécében benne van a munkamenet azonosítója. Szélsőséges esetben előfordulhat, hogy egy munkamenet azonosító csak egy személyhez tartozik.

Amennyiben munkamenet azonosító nem áll rendelkezésre vagy a nyomkövetéshez használt süti vagy a localStorage-ben tárolt adat megsérül vagy elérhetetlenné válik, a látogatóra jellemző paraméterek alapján érdemes az összekapcsolást elvégezni. Ha ez sem lehetséges, a böngésző és az eszközböngésző paraméterei segíthetik a látogatók követését és munkameneteinek összekapcsolását.

Bár a magyar nyelvben nincs megfelelő kifejezés a weboldalak összessége és egy konkrét weblap megkülönböztetésére, a továbbiakban a **webhely** kifejezést az azonos domain alatt elérhető oldalakra értem, a **weblap** kifejezést pedig egy konkrét oldalra. Egy webhely lehet például a WizzAir légitársaság honlapja és egy weblapra pedig a WizzAir webhelyének jegyfoglalási oldala. A webhelyek tehát weblapokból állnak, amelyeket gráf formában ábrázolhatunk, a csúcspontok maguk az weblapok és az élek a köztük lévő linkek. A 11. ábrán egy olyan honlap látható, amelyet fa struktúrában is lehet ábrázolni, azonban ez nem mindig teljesül, lehetnek olyan aloldalak is, amelyek egy másik aloldalra hivatkoznak.



11. ábra, Egy webhely weblapjainak gráf formában történő ábrázolása (saját szerkesztés)

A felhasználó a böngészés során egy honlap oldaltérképet különbözőképpen járhat be, ezt a bejárást nevezzük a **webhely bejárásának**, ami a webhelyen található első weboldal megtekintésétől annak elhagyásáig tart. Több kutatás is megerősítette, hogy a webhelyek bejárásának módja jellemző lehet az egyénre. (Shababi, Zarkesh, Adibi, & Shah, 1997)

Az következőkben a webhely bejárásának lehetséges fázisait elemzem a felhasználó azonosításának szintje szempontjából. Minden esetben értékelni fogom, hogy az adott fázisban

- ismert-e látogató kiléte, milyen mértékű a nyomonkövethetősége
- van-e lehetőség az eszközböngészők összekapcsolására
- ismert-e a felhasználó üzleti szempontból lényeges tulajdonsága.

A bejelentkezés nélküli böngészés

A megtekintett webhelyen nincs bejelentkezési lehetőség vagy a látogató nem jelentkezett be. Webáruházaknál gyakran alkalmazott technológia, hogy a látogató a webhely első weblapjának meglátogatása során kap egy munkamenet azonosítót és a virtuális kosarába pakolhatja a termékeket és csak akkor kell bejelentkeznie, ha tényleg meg is szeretné vásárolni a kívánt termékeket. Előnye, hogy a látogató nincs rákényszerítve, hogy már a terméklista böngészése előtt regisztráljon vagy belépjen, elég akkor, ha már fizetni szeretne. A munkamenet kiosztásának következményeképpen a webhely képes követni a felhasználóit.

A bejelentkezett felhasználó böngészése

A bejelentkezést követően az aktuális látogató kiléte elérhetővé és azonosíthatóvá válik a weblapok számára. Ez egyben egy logikai és adminisztratív védelemmel látja el a weblapot. (András & Péter, 2015)

Miután a látogató bejelentkezett a webhelyre, a nyomkövetése egyszerűvé válik, ugyanis a munkamenetben megtalálható azonosítója segítségével a rendszerben tárolt adatok közül egyszerűen kiválasztható a felhasználóhoz köthető adat, amelyet a regisztráció folyamata során előzőleg megadott magáról.

Ha a látogató különböző eszközökről is bejelentkezik, akkor az eszközböngészők egymáshoz kapcsolása is egyszerű, hiszen a lementett adatokból csak le kell szűrni a felhasználóhoz kapcsolható különböző eszközböngészőket.

A kijelentkezés utáni böngészés

A kijelentkezést követően a munkamenet azonosító és annak tartalma a szerveren törlődik, így már nem elérhető az látogatóhoz kapcsolható azonosító. Azonban nincs garantálva, hogy a felhasználó valóban ki lesz jelentkeztetve és a munkamenete törölve lesz, mint ahogyan 2013. februárjában egy hacker bebizonyította, hogy a Facebook nem először a kijelentkezés után is adatokat gyűjt a felhasználóiról. (Hobson, 2013)

Ha a kijelentkezés valóban megszünteti a munkamenetet, szintén nem túl bonyolult feladat a látogató későbbi életútját a szoftver és hardverkörnyezete alapján hozzákapcsolni, mivel ebben az esetben igen kis időtartamról beszélünk, ezért élhetünk azzal a feltételezéssel, hogy a látogatóról elmentett adatok jelentős része nem módosult.

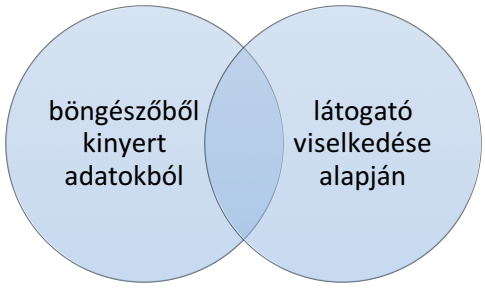
Egyazon bejelentkezési azonosítóhoz tartozó különböző eszközböngészők összekapcsolása a bejelentkezési azonosító miatt lehetséges, ez az információ kijelentkezést követően is megmarad.

3.4. Látogatók azonosításának és követésének módszerei

Az irodalom az egyének beazonosításra két módszert említ:

- **Felülről lefelé:** A teljes emberi populációból kiindulva szűkíti a kört addig, amíg meg nem találjuk a keresett egyént (Eckersley, 2013)
- **Egyén a tömegben**: Az emberek egyedi mobilitási képességein és viselkedésén alapuló eljárás (D. Blondel, A. Hidalgo, Verleysen, & de Montjoye, 2013)

A látogatókról kinyerhető adatok tárgyalását követően, belőlük építkezve lehetséges a látogatók által használt eszközböngészők összekapcsolása vagy a felhasználók profilozása. A felhasználók azonosítására és az eszközböngészők összekapcsolására több lehetőség nyílik, ezeket az alábbi csoportokba lehet rendezni:



12. ábra, Azonosítás módszereinek egymáshoz való viszonya (saját szerkesztés)

- **a böngésző által szolgáltatott adatokból kinyerhető információ segítségével:** a böngésző által küldött adatokból megismerhető a látogató által az aktuális munkamenethez használt eszközböngésző szoftver és hardver környezete. Ezzel csak az eszközböngészőt lehet beazonosítani, amivel az a probléma, hogy egy látogató több eszközböngészőt használhat és több felhasználó is használhat egy eszközböngészőt. (N:N kapcsolat) Így ez a módszer nem jelent tökéletes megoldást a problémára, azonban jelentősen lehet szűkíteni az eredményhalmazt.
- **viselkedés alapú:** a látogató online viselkedése meghatározhatja az egyént. Minél több információnk van a felhasználó böngészési szokásairól, annál pontosabban beazonosítható az egyén.
- **kombinált módszer:** az előbbi módszerek kombinálásával elérhető azonosítás

3.4.1. A böngésző által szolgáltatott adatokból kinyert információ segítségével

Előzőleg már megismertük, hogy milyen adatokat milyen módszerrel lehet kinyerni a látogató böngészőjéből. Fontos kiemelni, hogy egy hardver beazonosítása és követése nem jelent egyet a böngésző felhasználó azonosításával, mivel egy személynek több különböző eszköze is lehet, vagy az is előfordulhat, hogy más használja egy másik személy számára. Az alábbi táblázatban összefoglaltam, hogy az egyes azonosítási módszerek mit azonosítanak. Ennek megfelelő ez lehet

- **eszköz**: az internetezésre használt hardver és rajta lévő szoftverek (pl: mobiltelefon Android operációs rendszerrel vagy laptop Windows 7 operációs rendszerrel stb.)
- **böngésző**: az internetezéshez használt böngésző alkalmazás (Opera vagy Firefox stb.)
- **személy**: a hardvert használó személyre jellemző

A következőkben azt vizsgálom, hogy a megszerzett adatok segítségével hogyan lehet azonosítani, illetve követni a látogatókat, majd az alább látható táblázatban összefoglaltam, hogy a felsorolt azonosítási módszerek közül mely az, amely az eszközböngészőt azonosítja, meg az, amely az eszközt és melyik az, amely a felhasználót azonosítja.

Fontos különbséget tenni az eszközböngésző és az eszköz azonosítása között, ugyanis előfordulhat, hogy valaki kölcsönadja a gépét egy másik személy számára, ezért, ha az egyik módszerrel beazonosítottuk az eszközböngészőt, még nem biztos, hogy azzal beazonosítottuk a felhasználót is.

Azonosítási módszer	Mit azonosít? (eszközböngésző, eszköz, felhasználó)
Lokáció alapú	eszköz
Szoftver- és hardverkörnyezet	
Képernyő szélessége, magassága és bitmélysége	eszköz
Operációs rendszer típusa és verziószáma	eszköz
Böngésző típusa és verziószáma	eszköz
Eszköz típusa és verziószáma	eszköz
HTTP fejléc	eszközböngésző
HTML5 & CSS3 képességek	eszközböngésző
Cache alapú	eszközböngésző
Cookie, localStorage vagy egyéb WebStorage alapú	eszközböngésző
Telepített betűtípusok	eszköz
Viselkedés alapú	
meglátogatott webhelyek alapján	felhasználó
webhely bejárásának módja alapján	felhasználó
billentyűzethasználat alapján	felhasználó
egérmozgás követés alapján	felhasználó
Közösségi hálózatok által	felhasználó
Egyéb harmadik féltől hozzáférhető adat	felhasználó

5. táblázat: azonosításra használható paraméterek és azonosításuk tárgya

Az alábbiakban bemutatom a fontosabb böngésző által kínált nyomkövetésre használható adat típusokat. Az adatok egy részét Javascript segítségével a DOM-ból lehet kiolvasni, a másik részét pedig a HTTP fejlécből kiolvasni.

Lokáció alapú nyomkövetés

A látogató aktuális földrajzi pozíciójának meghatározására több módszer létezik. Egy hozzávetőleges földrajzi pozíciót az IP címből vissza lehet fejteni. Az IP címek kiosztása nem hierarchikus, azaz két egymás mellett IP cím nem jelenti azt, hogy földrajzilag is közel helyezkednek el egymáshoz. Ez a telefonszámok esetében igaz, ugyanis az ugyanabban a körzetben található kiosztott telefonszámok csak az adott körzeten belül fordulhatnak elő. Azonban léteznek olyan elérhető adatbázisok, melyek összerendelik a kiosztott IP címeket a földrajzi pozíciójukkal. Ebből az adatbázisból létezik ingyenes tartomány/város szintű, de a fizetős verzióban megtalálható akár az utca szintű felbontás is.

Az IP címek visszafejtéséhez az adatbázisban vissza kell keresni a látogató IP címét és az visszaadja annak földrajzi helyét a megfelelő pontossággal. Természetesen, ha a látogató valamilyen mobil eszközről éri az Internetet, akkor a módszer nem lesz pontos.

Sokkal pontosabb eredményt ad vissza a HTML5 GeoLocation szolgáltatása. Ez a HTML5 megjelenése előtt a Google Gears-ben[19] kísérleti jelleggel indult el, majd bekerült a HTML5 szabványba. Azért jóval pontosabb, mert ez a módszer nem az IP címből fejti vissza a földrajzi pozíciót, hanem a böngészéshez használt eszköz valamely földrajzi pozíciójának megállapítására alkalmas részegységétől kérdezi le az eszköz helyzetét. Egy mobil eszközben általában több olyan eszköz is megtalálható, amely segítségével megállapítható annak helyzete:

- **GPS/GLONASS:** akár méteres vagy deciméteres pontosság is elérhető
- **Wifi/Bluetooth/NFC:** különböző adatátviteli technológiák segítségével, a csatlakoztatott állomások pozíciójának ismeretében lehetséges a látogatáshoz használt eszközböngésző körülbelüli pozíciójának megállapítása
- **rádiótelefon cellainformáció:** amennyiben mobiltelefonról beszélünk, a hálózatra kapcsolt eszköz a 3 legközelebbi adótornyok pozíciójának ismeretében háromszögelés módszerével ki tudja számítani annak megközelítőleg pontos földrajzi pozícióját.

[19] A Google által 2007. május 31-én kiadott kiegészítő böngészőkhöz, a HTML5 megjelenése előtt RIAk (Rich Internet Application) számára biztosította a HTML5-ben megjelent lehetőségek egy részét (pl: GeoLocation, localStorage, SQLite, Worker, application cache)

A fenti listából a GPS/GLONASS adja vissza a legpontosabb eredményt, azonban egy mobil készülékben nem feltétlenül van mindig bekapcsolva, valamint felhasználói hozzájárulás is szükséges. Városi környezetben, ahol a mobiltelefon cellák sűrűn helyezkednek el, a háromszögelés módszerével is 10-20 méteres pontosság érhető el, akár fedett területen belül is. Egy GPS képes mobiltelefon esetében mindig a legpontosabb elérhető eszközt választja ki a helyzet meghatározásához.

A GeoLocation API kétféle lehetősége kínál:

- aktuális helyzet lekérdezése
- felhasználó helyzetének követése

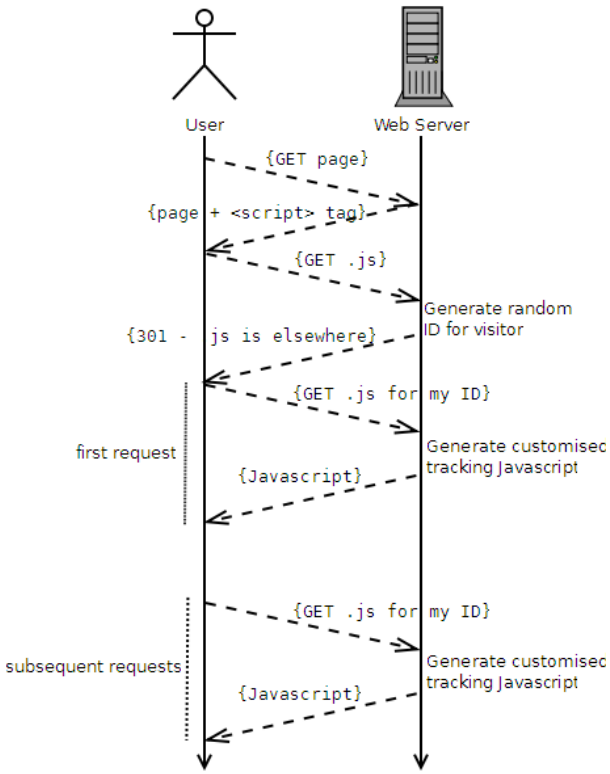
Cache alapú nyomkövetés

A sütik használatának hátránya, hogy könnyen törölhető a kliens eszközböngészőjéből. Nyomkövető sütik használatához pedig a felhasználó beleegyezését kell kérni. (Alexander, 2012) Azonban léteznek alternatív módszerek is a felhasználók követésére, egyik ilyen módszer a cache alapú azonosítás. A működés elve:

1. A felhasználó meglátogatja a webhelyet első alkalommal.
2. A letöltött weblap tartalmaz egy linket, amely egy külső JS file-ra mutat. A JS file nevét a szerver generálja és a szerveren az eszközböngészőhöz társítva elmentésre kerül. A JS file-t nem szabad cache-elni, mert a módszer előfeltétele, hogy minden alkalommal a böngésző leellenőrizze a JS file-t.
3. Böngészés közben a böngésző megkísérli letölteni a JavaScript file-t, amelyre a szerver HTTP 301-es hibaüzenetet küld vissza. A HTTP 301-es hibaüzenet azt jelenti, hogy a kért erőforrást áthelyezték ("Moved Permanently") és a HTTP válaszban benne lesz az erőforrás új helye.
4. A böngésző az új helyről letölti a JS file-t, amely képes betölteni eredeti funkcióját.

Ezzel az egyszerű trükkel nem szükséges a munkamenet sütik használata. A szerver a már előzőleg kiosztott azonosítót minden oldalletöltéskor meg fogja kapni, amikor a böngésző megpróbálja letölteni az oldal működéséhez szükséges állományt. Maga az eszközböngésző azonosító a lekérdezni kívánt file neve lesz. Ez a módszer a cookie-s azonosításhoz képest

- sokkal sebezhetőbb, mivel, ha a felhasználó kitörli a cache-t vagy Ctrl-F5-tel (force refresh) frissít az oldalra néhány böngésző esetében új nyomkövető azonosítót generál a szerver. Inkognitó módban nem használható.
- sokkal kevésbé blokkolható, egyelőre nem ismerik ezt a technikát a privacy filterek.
- a cookie-kal ellentétben ezt a módszert még nem szabályozzák a törvények.



13. ábra, Cache alapú azonosítás

A cookie, localStorage vagy egyéb WebStorage alapú nyomkövetés

A legelterjedtebb módszer a felhasználók követésére a süti alapú nyomkövetés. A működése nagyon egyszerű:

- első látogatás alkalmával a szerver egy egyedi kódot küld a kliens számára, amelyet az eltárol egy sütiben
- az oldalletöltések során a süti minden HTTP fejlécben el lesz küldve a szerver számára, ahol a fogadó alkalmazás az adatbázisból kikeresheti, hogy melyik felhasználóhoz tartozik az adott azonosító.

A localStorage, sessionStorage és más WebStorage alapú nyomkövetés esetében is ugyanígy működik, a kliens oldali azonosító tárolásában különböznek egyedül.

A telepített betűtípusok alapján történő nyomkövetés

Korábban már bemutatott telepített betűtípus lista lekérdezésének feltétele, hogy a kliens böngészőjére fel legyen telepítve Flash Player (ugyanis a JavaScripttel ellentétben a Flash képes lekérdezni az eszközre telepített betűtípusok listáját). Működése:

- A weboldal letöltődésekor az onDOMReady eseményt kiszolgáló függvényből kell meghívni a Flash modult, amely lekérdezi az eszközre telepített betűtípusokat (Ali, 2010)
- A művelet végeztével a Flash meghív egy JavaScript függvényt, amelynek átadja a betűtípusok listáját
- A kapott listát a JavaScript egy AJAX lekérdezéssel elküldi a szerver számára

Mivel ez a lista csak betűtípusok telepítése során változik (bővül) ezért rövid időtávon alkalmas a felhasználók követésére. Nagy előnye, hogy nem eszközböngészőt, hanem eszközt azonosít. Hátránya, hogy nem működik Apple termékeken, Flash modul használata nélkül vagy Flash blokkoló használata esetén. Az Adobe bevallása szerint a Flash Player elterjedése 99%-os a PC-ken. (Adobe, 2011)

Egy operációs rendszerre telepített programokkal több az adott alkalmazásra jellemző betűtípus is települhet, ez szintén azonosíthatja az eszközt. A betűtípus lista lekérdezésének feltétele, hogy a kliens böngészőjére fel legyen telepítve Flash Player (JavaScripttel nem lehet lekérdezni a telepített betűtípusok listáját).

3.4.2. A viselkedés alapú nyomkövetés

A viselkedés alapú nyomkövetési módszerek közös jellemzője, hogy a felhasználót képesek azonosítani. Működési elve azon alapszik, hogy a látogató a viselkedésében hordozza az egyediséget, amely a megfelelő módszerekkel azonosíthatóvá válik. A kutatásaim során az alábbi módszerekre bukkantam:

Meglátogatott weblapok alapján történő nyomkövetés

Az internethasználókra, mint egyénekre jellemző, hogy azokat a weblapokat látogatják meg, amelyek őket érdekli. Az egyéneket az ilyen fajta internet használati szokásuk szerint is lehet csoportosítani. Azonban a felhasználók különböző eszközökről történő tartalomfogyasztása akár jelentős mértékben eltérhet egymástól. A látogatók böngészési előzményeiből az alábbi jellemzők lehetnek alkalmasak a felhasználó nyomkövetésére:

- Az egyén által meglátogatott webhelyek listája
- Az egyén által meglátogatott webhelyek illetve weboldalak látogatásának sorrendje
- Milyen eszközökről és honnan látogatja meg az egyén az adott webhelyet/weboldalt

A látogatók által meglátogatott weblapok teljes listája a böngészőből kiolvasva biztonsági okok miatt nem lehetséges. A közösségi hálózatok - mint a Facebook - amelyeknek sok weboldalba be van ágyazva egy komponense, minden látogatás alkalmával elküldi a Facebook számára a meglátogatott weboldal címét, így pontosabb kép alkotható a látogatók böngészési szokásairól. Minél több weboldalba van beágyazva egy közösségi oldal komponense, annál pontosabb képet kaphat a közösségi oldal a felhasználói által látogatott weboldalak köréről.

A webhely bejárásának módja alapján

A meglátogatott webhelyek bejárásának módja szintén jellemző lehet az egyénre. A különféle weboldal bejárási módok csoportosíthatóak és szintén használható az egyén nyomkövetésénél bizonytalanság csökkentésére.

Az egérmozgás és billentyűzethasználat alapján

A biometrikus azonosítás azon alapszik, hogy minden egyénnek saját használati szokásai vannak, az emberi lét és viselkedés paramétereit használja ki. A billentyűzethasználat esetén kijelenthető, hogy a különféle emberek gépelési sebessége között van különbség. Elemzések szerint 300 karakter begépelését követően már elegendő adat gyűlik össze ahhoz, hogy valakinek elkészítsék a profilját. A módszer az esetek 96%-ban működik helyesen. (Peter, David, David, Warren, & Jonathan, 2005) Az egérmozgást követő kísérletek esetében 20 kattintás elegendő volt ahhoz, hogy az algoritmus csak 1,3%-ban tévedjen. (Shivani, 2004) (Nan, Aaron, & Haining)

A közösségi hálózatok által történő követés

A közösségi hálózatokról lekérhető adatok minden esetben egyértelműen azonosítják a látogatókat, azonban ehhez szükséges a felhasználók hozzájárulása, amelyet a felhasználók elenyésző százaléka ad meg, még akkor is, ha egy megbízható oldalról van szó. A kísérletemben nyereményjátékkal próbáltam ösztönözni a felhasználókat privát adatait megadására. Kutatásom során azt tapasztaltam, hogy egy

a felhasználók által jól ismert oldal esetén is csak kevesen engednek hozzáférést az adataikhoz.

Egyéb harmadik féltől hozzáférhető adat alapján

A közösségi hálózatokon kívül más harmadik fél is használható a látogatók azonosítására. Amíg a közösségi hálózatoknak ez az egyik szolgáltatása, az ebbe a kategóriába kerültek azok az oldalak, amelyek a látogatókról valamilyen információt tárolnak és potenciálisan hozzásegíthetnek az egyének beazonosításához:

- **youhavedownloaded.com** egy olyan oldal, amely arról vezet naplót, hogy milyen IP címről mikor töltöttek tartalmakat. Az oldal jelenleg nem elérhető, de statisztikájuk szerint 55 millió felhasználóról tárolnak közel 115 TB-nyi adatot.
- **Shodan**: gonosz Google-nek is hívják, IP címre lehet rákeresni és az oldal visszaadja, hogy az adott IP címen milyen portok vannak nyitva, vagy milyen IP kamerák elérhetőek
- **Whatismyip.com** oldalról kinyerhető, hogy a látogató a webhely eléréséhez proxy-t használt-e, valamint, hogy az adott IP cím tiltólistán van-e

3.4.3. A látogató azonosítása és követése

Dolgozatomban a látogató **követésének** nevezem azt a folyamatot, amikor az eszközböngészők által létrehozott munkamenetek közötti kapcsolat kimutatható. A látogató **azonosításán** értem azt a folyamatot, amikor a látogató önkéntesen megadott adatai alapján, az internetezéshez használt eszközének szoftver és hardverkörnyezetének alapján vagy a látogató viselkedésének jellemzőiből a látogatóra egyértelműen következtetni lehet.

Egy látogató követéséhez szükséges ismerni az általa használt eszközböngészőket. Alapfeltevésként fogadjuk el, hogy a felhasználók több internetezésre alkalmas készülékkel rendelkezhet. Ezeket a készülékeket más személyek is használhatják, de abból indulok ki, hogy az esetek nagy részében egy látogató a saját készülékét használja.

Feltételezhetjük, hogy a felhasználók zöme legalább nem változtatja meg szándékosan az általa használt eszköz szoftverkörnyezetét, hogy az a nyomkövető alkalmazást megtévessze. Szintén nem életszerű, hogy a látogató a munkamenete ideje alatt a detektálható paramétereinek jelentős részét megváltoztatja.

A látogató által használt különböző eszközböngészők egymáshoz kapcsolása egy vagy több eszközfüggetlen paraméter segítségével jöhet létre. Az összekapcsolás során a potenciális eszközböngészők körének szűkítéséhez az alábbi szabályok is felhasználhatóak:

- a látogatók valószínűleg nem rendelkeznek több vezetékes internetkapcsolattal
- a felhasználók helyváltoztatásának korlátossága
- a nem mobil eszközök esetében a munkamenetek alatt a felhasználó földrajzi pozíciója nem változik

Az alábbiakban az összekapcsolásra alkalmas paramétereket veszem sorra és kiértékelem:

- rövid időintervallumon belül egyazon **IP cím** használata/azonos hálózat használata:
    - egy látogató egy adott hálózatból több eszközböngészőt is használhat, jó példa erre az az eset, amikor a felhasználó a laptopját és a mobiltelefonját is használja az otthoni hálózatból
    - az otthoni internet előfizetéssel igen gyakran 1 db IPv4-es IP címmel több eszköz is képes elérni a hálózatot a NAT/PAT technológiának köszönhetően
- nem mobil eszközök esetében azonos **földrajzi pozíciónál**
    - **vezetékes kapcsolódás esetén** a pozíció meghatározás nem pontos, csak város szintű feloldást eredményez, így az ott megállapítottak alkalmazhatóak erre az esetre is
    - **vezetéknélküli kapcsolódás** esetén pozíciót lekérdezni (A-GPS, GPS, GLONASS, Wifi) képes eszközzel felszerelt egységgel rendelkező hardver esetén, az **IP címből** visszafejtve a pontosság kérdéses lehet, ennek mértékétől függően a közel ugyanabból a pozícióból érkező lekérdezések szintén potenciális alternatív eszközböngészőt jelenthetnek
- **felhasználó jellemzői** alapján
    - **billentyűzet, egér, tapipad** és **érintőképernyő használat**: az internet eléréséhez használt fejlett beviteli eszközök vezérlése a látogatóra jellemző, akárcsak az aláírás. Ez a biometrikus azonosítás egyik fajtája, a viselkedés szerinti azonosítási módszerek közé tartozik. A módszer

nem biztosítja ugyanazt a szintet, mint pl. egy retina szkennelés, viszont a használatához nem szükséges drága hardvert beszerezni. A módszer azon alapszik, hogy az egyének gépelési stílusa egyedi. A billentyűzetfigyelés során olyan alapvető tulajdonságokat mérnek, mint a sebesség, a lenyomás ereje és a nyomvatartás ideje. Az azonosításhoz csupán egy 300 karakterből álló mintára van szükség, amiből kinyerhető az egyénre jellemző információ és a későbbiekben referenciaként használható. A módszer az esetek 96%-ban működött helyesen. (Peter, David, David, Warren, & Jonathan, 2005) Míg az egérmozgást követő kísérletek esetében 20 kattintás elegendő volt ahhoz, hogy az algoritmus csak 1,3%-ban tévedjen. (Shivani, 2004) (Nan, Aaron, & Haining) A billentyűzet és egérmozgás figyelése (akár a telepített betűtípusok listájának lekérdezése is) azonban az egérmozgás során rengeteg adat keletkezik, mivel a minél nagyobb a minta, annál pontosabb lesz az elemzés. Az adatokat célszerű tömörített formában átküldeni a szerveroldalra és a már meglévő minták között hasonlóakat keresni.

- **hasonló weboldalak meglátogatása**: a látogató által meglátogatott weblapok halmaza az egyénre jellemző, így ez alapján be lehet azonosítani a felhasználót és ezzel a különböző eszközböngészőket egymáshoz lehet kapcsolni

Fontos kiemelni, hogy a felhasználó és a jellemzők összekapcsolása feltételes. Ha egy esemény többször is bekövetkezik, akkor annak a valószínűsége, hogy a két eszközböngésző egy személyhez köthető megnő. (pl: ha minden hétköznap egy adott ISP-hez csatlakozva ugyanabból a hálózatból éri el egy személy az internetet mobiltelefonnal és laptop használatával) Célszerű a feltételezett alternatív eszközböngészőket elmenteni, majd ahogy idővel egyre több adat gyűlik össze a látogatóról felülvizsgálni annak elemeit.

3.5. Összegzés

A webes böngészési alapjainak bemutatását követően az internetezéshez használt eszköz szoftver és hardverkörnyezetét vettem górcső alá, majd az internetező személy online tanúsított viselkedését jellemző és hozzáférhető paramétereit

vizsgáltam meg, amelyek felhasználásával a személye beazonosíthatóvá válhat. Tehát a téma megértéséhez szükséges ismeretek rendszerezett leírását követően, megállapítottam a böngésző segítségével elérhető adatok lehetséges halmazát, majd ezen adatok elérésének módját és végül az egy domain alól elérhető weboldalak számára összegyűjthető adatokból levezettem a látogatók követésének módszerét.

Az internetet használók több eszközt és azon több különböző eszközböngészőt is használhatnak. A látogató online viselkedésének átfogó elemzéséhez ezen eszközböngészők összekapcsolása szükséges. Az eszközböngészők összekapcsolhatóságának valószínűsége annál magasabb, minél szélesebb rálátásunk van a látogató online viselkedésére, cselekedeteire, időben és adatminőségben egyaránt.

Az üzlet szempontjából fontos jellemzők szerint csoportosított felhasználók számára küldött reklámok hatékonysága jóval magasabb, mint a hagyományos reklámoké. A cél határozza meg az azonosítás szintjét: Az üzlet számára nem fontos, hogy a látogató pontos kilétét ismerjük, az is elegendő, ha csak néhány kiemelt tulajdonságát ismerjük. Az ebben a fejezetben leírt technológiai modellek és technikák a következőkben ismertetett kutatásaim alapjául szolgáltak.

4. AZ ESZKÖZBÖNGÉSZŐKRŐL BEGYŰJTHETŐ ADATOK JELLEMZŐI

Ebben a fejezetben leírom a kutatásaimhoz nélkülözhetetlen adatgyűjtési fázist, valamint a rá épülő első kutatásomat, az egy domain alól elérhető weblapok látogatóinak eszközböngészőjéről begyűjthető adatokból kinyerhető információk elemzését mutatom be.

4.1. A kutatás bemutatása

A látogatók több különböző eszközt használnak internetezésre (pl: mobiltelefon, tablet vagy laptop) melyekről a tartalomfogyasztási és eszközhasználati szokásaik nagymértékben eltérhetnek. A látogatók online szokásainak teljes felméréséhez, nem elég egy eszközről gyűjtött adatok elemzése, hanem szükséges a különböző eszközök és böngészők összekapcsolása is.

Amíg a látogató internetezésre használt eszközének paramétereinek lekérdezése egyszerű, maguknak a látogatók beazonosítása már nem triviális. Kutatásomban annak járok utána, hogy miként használható fel a látogató által használt böngészők és eszközökről lekérdezhető adatok a látogató azonosítására. Utánajárok, hogy milyen látogatóra jellemző adatok – például a billentyűzet és egérhasználat - kérdezhető le a böngésző segítségével és ezek milyen mértékben alkalmasak a látogató azonosítására. Elemzem, hogy a látogató eszközböngészőjéből hozzáférhető adatok milyen mértékben járulnak hozzá a látogató beazonosításához.

Egy oldalletöltés alkalmával néhány kilobyte-nyi adat hozzáférhető a meglátogatott webhely számára, feltételezésem szerint ebből adatbányászati módszerekkel értékes – akár személyiségi jogokat sértő – információk nyerhetőek ki.

4.2. Az adatgyűjtés

Az adatgyűjtési fázisban a dolgozatban a 3 kutatási célhoz szükséges minta begyűjtésének folyamatát és tervezésének lépéseit írom le. A megfelelő minőségű minta biztosításához az alábbi pontokban fogalmaztam meg az adatgyűjtő alkalmazás tulajdonságait:

- a **hozzáférhető adatok legszélesebb körét** képes legyen elmenteni
- **alacsony szintű** hozzáférést biztosítson az **összegyűjtött adatokhoz**

- az **ismert blokkoló alkalmazások** (Ghostery stb.) ne legyenek képesek megakadályozni a működését
- adatgyűjtő alkalmazás működése közben az **átlagos felhasználó látogató ne tudja**, hogy adatait kezelem

Megvizsgáltam az interneten közzétett adatgyűjtő alkalmazásokat, azonban nem találtam olyat, amely az általam felállított feltételek mindegyikét kielégítette volna. Emiatt saját alkalmazás fejlesztése mellett döntöttem.

A fejlesztés megkezdése előtt számba vettem a felhasználó azonosításának módszereit, amelyeket alapvetően két csoportba lehet elkülöníteni:

- A látogató a saját **adatai önkéntes megadásával** engedi saját magát azonosítani és emiatt követni
- A látogató pusztán az internetezéshez használt **hardver és szoftverkörnyezetének**, valamint a saját **viselkedésének jellemzőivel** teszi lehetővé saját maga azonosítását. Az egyén az érdeklődési körének megfelelő weboldalakat látogatja meg, így a meglátogatott weblapok listájából következtetni lehet a látogató személyére (Jia-Ching, Chu-Yu, & Vincent, 2012)

Amennyiben a látogatók egyedi azonosító adataik megadásával felfedik magukat, online beazonosításuk triviálissá válik, emiatt a kutatásom célpontjai a második kategóriába eső látogatók csoportja volt.

A már korábban definiált adatgyűjtő weblaptípus kategóriák szerint a saját fejlesztésű adatgyűjtő alkalmazásom az egyszerű weblapok/adatgyűjtők csoportjába tartozik, amellyel a látogatók online viselkedési jellemzőinek kis szelete érhető el, míg a kiterjedt hálózattal rendelkező weblapok a felhasználó online viselkedéséről átfogó képet kaphatnak.

A lementett adatokat időbeliségük szerint az alábbiak szerint lehet elemezni:

- a lementett **rekordok** elemzése: egy elmentett rekord a látogató böngészésének egy időpillanatát rögzíti
    - a pillanatfelvételek elemzésével az oldal **látogatóinak összességéről** is nyerhető információ

- eszközböngészőhöz tartozó **munkamenetek** elemzése: az eszközböngészők számára kiosztott egyedi azonosítók biztosítják az eszközböngészőről megkezdett munkamenetek elemzését
    - lehetőség nyílik a látogató eszközböngészője paramétereinek időbeli vizsgálatára
    - mobil eszköz esetén a látogató által meglátogatott helyek kideríthetőek lehetnek
    - a meglátogatott weblap bejárásának módja elemezhetővé válik
- az egy látogató által használt **összekapcsolt eszközböngészőkön** átívelő elemzési módszer használatával: a látogató által használt valamennyi online viselkedési variáció megismerhető
    - elérhetővé válnak a látogató által használt eszközböngészők munkamenetei

A kutatásom során az egy oldalba beépített adatgyűjtő alkalmazásom pusztán a mindenki számára hozzáférhető adatokra támaszkodik, az elemzési fázisban külön elemeztem a rekordokat és a látogatók munkameneteit.

A felhasználó azonosítása az internetezésre használt eszköz és böngésző paramétereinek azonosításával, illetve a felhasználó eszközhasználati módjának azonosításával lehetséges.

A böngésző és az internetezésre használt eszköz hardver és szoftver paramétereinek lekérdezésével magát az eszközt lehet azonosítani. A paraméterek többsége rendelkezésre áll, pusztán csak ki kell olvasni a kliens vagy a szerveroldalon. A közvetlenül nem hozzáférhető többi adatot vagy egy harmadik féltől lehet lekérdezni vagy valamilyen kiegészítő eszköz segítségével érhetőek el.

Az eszköz használatának módja jellemző a felhasználóra. Az eszköz használatának módját egyrészt az input eszközökön keresztül lehet detektálni. Ez jellemzően az egér vagy az érintőképernyő, amelynek a használata az aláíráshoz hasonlóan az egyénre jellemző, hiszen ugyanolyan finom motorikus mozgások szükségesek hozzá. Szintén jellemző lehet az egyénre az általa meglátogatott weboldalak köre vagy az érdeklődési körébe tartozó weboldalak megtekintése.

4.2.1. A célcsoport

A kutatás tervezésekor a célom az volt, hogy a lehető legtöbb felhasználót érjem el. Az BCE e-learning rendszerének használata ebből a szempontból ideális, mivel ezzel azt a hallgatók közel fele elérhető.

Az Egyetem e-learning rendszerébe illesztett adatgyűjtő alkalmazást csak a rendszer felhasználói érhetik el. Belépni csak a felhasználói azonosítóval és jelszóval lehetséges. Minden látogató csak ahhoz a kurzushoz férhet hozzá, amihez hozzárendelte a kurzus adminisztrátora. A belépésre jogosultak a BCE hallgatói, tanárai és kisebb számban az e-learning rendszer karbantartói. A megfigyelési egységek az egyének. A diákok túlnyomórészt az X és a Y generáció tagjai:

- **X generáció** tagjai 1960 és 1980 között születettek. Jellemzőik a megbízhatóság, elmélyült szakmai igényesség, magas motiváció, kooperativitás és karrierizmus. A digitális technológiával már fiatalon megismerkedtek, de a következő generációkhoz képest ezen a területen alulmaradnak.
- **Y generáció** tagjai 1980 és 2000 között születtek, főbb jellemzőik közé tartozik az elmélyült tudás iránti igény és a munkára, tanulásra való motiváltság gyengülése (Regina, 2012)

A kutatás végeztével a kapott adathalmaz sem a magyar populációra, sem az egyetemi hallgatókra nézve nem tekinthető reprezentatívnak.

4.2.2. Az adatgyűjtő alkalmazás

A Budapesti Corvinus Egyetem e-learning rendszerébe beépített, általam fejlesztett alkalmazás 2012. május 3 és 2012. május 22 között gyűjtötte a látogatók adatait. Az alkalmazás működésének leírása megtalálható a „Adatgyűjtő alkalmazás fejezetben”.

Mivel a kutatás során személyes adatokat is kezelek, ezért indulás előtt az alkalmazást auditálásnak kellett alávetnem, valamint meg kellett szereznem az intézeti adatvédelmi biztos engedélyét és az e-Learning és Oktató- és Szolgáltató Központ vezetőjének hozzájárulását az alkalmazásom az e-learning rendszerbe való beépítéséhez. A kutatás tervezésénél a kérvények és engedélyek beszerzése hozzávetőlegesen 3 hetet vett igénybe.

Az alkalmazás elkészültét követően annak teljes forráskódját el kellett küldenem az Intézeti Adatvédelmi felelős által kijelölt bizottságnak felülvizsgálatra. Ezt követően történhetett meg az alkalmazás beépítése.

A tesztelési fázis alatt az alkalmazást különböző böngészőkben, más és más operációs rendszerek alól, valamint különböző eszközökkel (számítógép, táblagép, mobiltelefon, okostelefon) teszteltem, hogy megbizonyosodjak, az a legtöbb esetben az elvártaknak megfelelően működik. A tesztfázis sok különféle hiba kiszűrésére alkalmas. (pl: mindenféle eszközön jól olvashatónak kell lennie a szövegeknek vagy a mobiltelefon kijelzője túl kicsi a többsoros szöveg megjelenítéséhez)

Az adatgyűjtő alkalmazás egyik fontos lépése volt a kinyerhető adatok relevanciájának és annak meghatározása, hogy az adatgyűjtés mekkora terhet ró a kliens és a szerveroldalra, hogy elkerüljem a felesleges adatok mentését. A becslés alapján adatbázist hoztam létre az adatok tárolására.

Az adatok kinyeréséhez szükséges hardver erőforrás azt mutatja, hogy a paraméterhez való hozzáférés vagy az adott paraméter kinyerése milyen mértékben terheli a kliens és/vagy a szerver erőforrásait és mennyit kell várni a kinyerni kívánt adatra:

- **alacsony:** alacsony mértékben terheli a hardvert egy érték kiolvasása vagy környezeti változóhoz való hozzáférés
- **közepes:** pl. egy script futtatása, amely képes felmérni a böngésző képességét, vagy egy DNS lekérdezés
- **magas:** pl. online API-tól való lekérdezés vagy a processzort erősen terhelő művelet
- **nagyon magas:** komoly statisztikai és adatbányászati műveletek esetén

Az adatok relevanciája azt jelenti, hogy azok milyen mértékben járulnak hozzá a látogató beazonosításához és követhetőségéhez:

- **alacsony:** kismértékben járulnak hozzá a felhasználó eszközböngészőjének követéséhez, a paraméterek értéke nagy valószínűséggel nem egyedik. Pl. nagyon sok böngésző támogatja a localStorage-et, azonban az összes HTML5-ös jellemző egyedi lehet
- **közepes:** 4-8 csoportra osztja a látogatókat, ezzel csökkenti a látogató kilétére vonatkozó bizonytalanságot

- **magas:** a tényező ismerete nagyban hozzájárul a felhasználó azonosításához, nagy valószínűséggel jellemzi az eszközböngészőt. Hátránya lehet, hogy viszonylag rövid időn belül változhat.

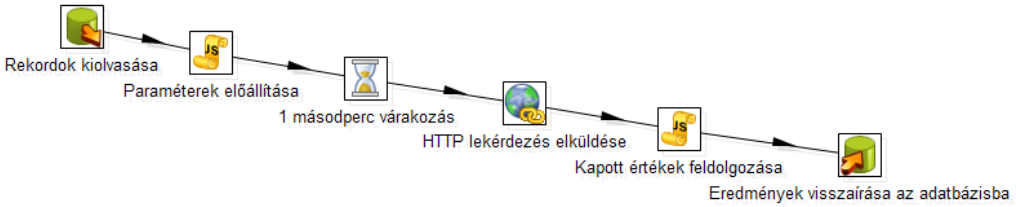
A változók részletes listája és az értékelésük teljes listája megtekinthető a 8.7.2 fejezetben. Néhány fontosabb megfigyelés:

- A HTML5 képességek böngészőre és azok verzióira jellemzőek, így azokat felesleges detektálni és lementeni.
- Az egérmozgás követése és leütött billentyűk elmentése nagyon leterhelheti a kliens oldalt és a hálózatot. A felhasználói élmény romlásától tartva, nem fejlesztettem bele az adatgyűjtő alkalmazásba.
- A localStorage kliens oldali adattároló technológia megbízhatóbb, mint a sütiben történő adattárolás. Így, ha a süti tartalma valamiért elvész vagy megsérül, az eszközböngésző azonosító a localStorage-ből még visszaállítható.
- Az inkognitó mód detektálására már nincs lehetőség, mert a hibát, amely révén hozzáférhető volt ez az információ a böngészők fejlesztői már kijavították.

4.2.3. A lementett adatok

Az alkalmazás által lementett adatok teljes listája a függelék 8.7.3 fejezetében tekinthető meg. Az adatgyűjtési időszak után következett az adatok előfeldolgozása az alábbiak figyelembevételével:

- A fizikailag a Corvinus Egyetemen található számítógépeket kivettem az elemzési mintából, mivel az ott található számítógépekről minden belépés alkalmával letörlődnek a sütik és szoftver és hardver kiépítettségük is hasonló.
- Az **IP címből** kinyerhető az **internet szolgáltató** vagy az internetezéshez használt **intézmény** neve, melynek értékét elmentettem a host mezőbe
- Azoknál a rekordoknál, ahol a látogatók nem adták hozzájárulásukat a földrajzi hely megosztásához, az **IP címeket** leképeztem **hosszúsági és szélességi fokokká**
- A Google Map API-jának segítségével a földrajzi pozíciókat **városra, utcára és házszámra** oldottam fel a 14. ábra*: a földrajzi pozíciók város, utca és házszámra történő feloldása Pentaho-val Google Maps API-n keresztül* látható módon



14. ábra: a földrajzi pozíciók város, utca és házszámra történő feloldása Pentaho-val Google Maps API-n keresztül

- Az adattábla **referrer** mezőjének értéke alapján kinyertem, hogy a látogató mely **tantárgyhoz** tartozó oldalt töltötte be
- A **mobil eszközökön** található **operációs rendszert** kézzel vittem fel az adatbázisba
- A **böngésző típusát** (browserFamily), az **operációs rendszer típusát** (osFamily) és **verzióját** (osVersion), valamint az **eszköz** (device) mezők értékét az UAParser PHP könyvtár segítségével a HTTPUserAgent mező értékéből kinyerve töltöttem ki

4.3. A bizonytalanság-csökkentő képesség mérőszáma

A dolgozatomban a látogatók által használt eszközböngészőkből kinyert adatok bizonytalanság-csökkentő képességének méréséhez Athanasios S. Voulodimos & Charalampos Z. Patrikakis által a „Quantifying privacy in terms of entropy for context aware services". (Voulodimos & Patrikakis , 2009) című cikkében közzétett keretrendszert, valamint a Peter Eckersley „A Primer on Information Theory and Privacy" (Eckersley, 2013) cikkében közzétett technikákat használom Az általuk használt keretrendszer Claude E. Shannon által 1949-ben publikált bizonytalanság számszerűsítésének matematikai módszerén alapszik. A bizonytalanság fokát Shannon entrópiában (S) fejezi ki. (Shannon, 1948) Egy X rendszer információ entrópiájának képlete:

$$S(X) = -K \sum_{j=1}^{M} p_j log(p_j)$$

Ahol:

- **K** különböző mértékegységek összehasonlításához használt konstans (dolgozatomban az értékét K = 1-nek választom)
- **M** a vizsgált rendszer alrendszereinek értékeinek/állapotainak számossága
- $\boldsymbol{p_j}$ a j-edik elem kiválasztásának valószínűsége ($ahol \sum_{j=1}^{M} p_j = 1$)

Megjegyzem, hogy az informatikai adatok entrópiájának mérésére kettes alapú logaritmust használok, tehát a képletekben szereplő logaritmus függvény minden esetben kettes alapú ($\log = \log_2$). A képlet csak az alábbi feltételezések és megkötések mellett használható:

- A látogatót jellemző paraméterek azonos valószínűséggel vehetik fel az értékeiket (monotonitás)
- Egy alrendszerekből álló rendszer entrópiája a diszjunkt alrendszerek[20] entrópiájának súlyozott összege. Tegyük fel, hogy S_c a következő elkülönülő alrendszerekből áll: S_a és S_b ($a = |S_a|/|S_c|$ és $b = |S_b|/|S_c|$), ekkor $H(S_c) = H(a, b) + aH(S_a) + bH(S_b)$ a rendszer entrópiája. (Voulodimos & Patrikakis , 2009) (rekurzivitás)

Az előbbiekben megismertetett entrópia modell használható a látogatókról összegyűjtött adatok minőségi jellemzésére, használatával összehasonlíthatóvá válnak a mintában szereplő változók információfelfedési képességei.

4.4. Az összegyűjtött adatok vizsgálata

A kutatásom exploratív, azaz feltáró jellegű, hipotéziseket nem definiáltam, ehelyett az alábbi, a kutatás irányát meghatározó kérdéseket és kérdésköröket fogalmaztam meg:

- Az eszközböngészőkből származó adatok segítségével, személyes adataik felhasználása nélkül lehetséges-e a látogatók beazonosítása és követése?
    - Milyen adatok vagy az adatok mely csoportja járul hozzá hatékonyan a felhasználók azonosításához és követéséhez a látogató által használt eszközökön keresztül?

[20] Egy rendszer alrendszereiben lévő elemek között nincs olyan, amely valamely más alrendszerébe is beleillik. A csoportosítás alapját képező tulajdonságok vagy tulajdonságok csoportja egyértelműen képes eldönteni a vizsgálandó elem alrendszerbeli tagságát.

- Egy weboldal megtekintésével elérhető-e elegendő mennyiségű adat a látogató személyéről vagy környezetéről ahhoz, hogy a személyes adatai nélkül azonosítani és követni lehessen?
    - Átlagosan milyen mennyiségű adat érhető el egy látogatóról egy weboldal megtekintése során?
    - Milyen módszerekkel lehetséges mérsékelni a weblapok számára elérhető adatokat?
    - A látogatóról megszerzett adatok milyen mértékben járulnak hozzá az azonosításához?
    - A hozzáférhető paraméterek milyen mértékben képesek eloszlatni a látogató beazonosításának bizonytalanságát?

4.4.1. Alapvető statisztikák

A kutatás adatgyűjtési fázisát követően az alábbi statisztikákat szereztem be. A Moodle-re vonatkozó adatokat a rendszerbe ágyazott Google Analytics-ből nyertem ki, amelyet a Moodle karbantartóitól az e-Learning Oktató- és Szolgáltatóközpont munkatársaitól kértem el.

- Látogatási statisztikák
    - 2013. március havi oldalletöltés: **kb. 134 000**
    - 2013. március havi egyéni látogató: **kb. 31 000**
- Moodle-ből nyert adatok:
    - **8 169** különböző felhasználó lépett be az elmúlt hónapban
    - **8 542** különböző felhasználó lépett be az elmúlt félévben
- Corvinus weblapról nyert információ[21]
    - hallgatói létszám: **17 879**
    - tanárok létszáma: **867**

Az adatgyűjtési fázisban **647 242** rekordot mentett el az alkalmazás az adatbázisba, amelyek jellemzői a 8.7.4 fejezetben tekinthetőek meg. Az egyes mezők értelmezése rendre:

- **lementett paraméter neve**

[21] http://web.uni-corvinus.hu/subpage_choice_control.php?org=2&id=8&UC=&subpage=&LNG=eng

- **kitöltött érték:** azon rekordok száma, amelyekben a kérdéses paraméter ki volt töltve (a kliens vagy a szerver oldalon a paraméter elérhető volt, az lementésre került az adattábla megfelelő oszlopába), zárójelben mögötte a százalékos érték található
- **különböző értékek:** ahány különböző rekord található az adatbázisban az adott paraméterre nézve
- **különböző értékek a kiosztott eszközböngészők arányában:** az aktuális paraméter és a kiosztott felhasználói azonosító számának hányadosa.

A táblázat fejléc alatti első sorában látható, hogy a két hét alatt **32 529** darab eszközböngésző azonosítót osztott ki az alkalmazás, ha ezt az értéket elosztjuk a belépett felhasználók számával, akkor minden egyes felhasználóra **3,98** eszközböngésző azonosító jut, tehát minden egyes felhasználó majdnem 4 különböző eszközböngészővel lépett be. Ez az érték ugyan kevesebb, mint a Cisco felmérésében szereplő érték, azonban esetemben a minta a Corvinus Egyetem polgárai, ez torzított sokaság. A felhasználók jelentős része feltehetőleg legalább egyszer használta a tantermi gépeket, melyek közös jellemzője, hogy fix IP-vel rendelkeznek, valamint közel hasonló kiépítettségűek, így a lementett paraméterek alapján nehéz különbséget találni köztük.

A lementett adatok jellemzőinek listája a 8.7.4 fejezetben tekinthető meg. Néhány fontosabb paraméter és értelmezése az alábbiakban olvasható:

- A mintában 2173 különböző HTTP user agent található, ez a kiosztott eszközböngészők 6,68%-a
- A telepített betűtípusok listájának lekérdezése az esetek 95,68%-ban elérhető volt.

Az alábbi paraméterekkel lehetséges az eszközböngészők legalább 5%-át beazonosítani:

- **IP cím:** az esetek nagy részében csak a munkamenet ideje alatt képes azonosítani az eszközböngészőt
- **HTTP user agent:** a böngészőről, az operációs rendszerről és a böngészéshez használt eszköz egy szövegbe való sűrítése
- **URL:** a felhasználó által megtekintett weblapok

- **Telepített betűtípusok:** közel 50%-os eredményt lehet elérni a telepített betűtípusok listájának lekérdezésével

A válaszok keresése közben arra a megállapításra jutottam, hogy pusztán az eszközböngészőkből származó adatokból a lementett mintámban nem tudok kimutatni kapcsolatot a személyhez nem köthető adatok és a látogatók személyes adatai között, a magukat szándékosan felfedni nem kívánó látogatók esetében. Ennek a kutatásnak a részleteit a dolgozatomban nem tartalmazza.

A fenti ok miatt a böngészők számára **hozzáférhető paraméterek bizonytalanság-csökkentő erejét** vizsgáltam, bemutatva az egy domain alól elérhető weboldalak és a közösségi oldalak látogatóikról elérhető adatok mennyiségét és minőségét.

4.4.2. A lementett változók bizonytalanság csökkentő ereje

A kutatás első lépéseként az elemzendő változókat fa struktúrába rendeztem majd külön-külön határoztam meg a bizonytalanság csökkentő erejüket. Az ábra megtekinthető a függelék 8.7.3 fejezetében. A számolás eredménye a 15. táblázatban látható.

Paraméter	Lementett értékek száma	Egyedi elemek száma	Bizonytalanság csökkentő tényező (ΔS) (bit)
Facebook azonosító	169	139	0,28
IP cím	32 529	14 443	1,17
DNT	32 529	2	13,98
Képernyő szélessége és magassága	32 529	509	5,99
HTTP User Agent	32 529	2 078	3,96
Böngésző típusa	32 529	7	12,18
Böngésző típusa és verziója		243	7,06
Operációs rendszer	32 529	27	10,23
Operációs rendszer típusa és verziója	32 529	101	8,33
Flash blokkolt-e	32 529	2	13,98
Flash verzió	32 529	104	8,28
Silverlight	32 529	26	10,28
Quicktime	32 529	39	9,70
Java	32 529	49	9,37
PDF Reader	32 529	2	13,98
Adobe Reader	32 529	90	8,49
Kapcsolat típusa	32 529	4	12,98
Földrajzi pozíció (IP címből visszafejtett)	32 527	180	7,49
GeoLocation	236	155	0,60
Telepített betűtípusok	32 529	8356	1,96

15. táblázat: Lementett változók bizonytalanság csökkentő ereje

Összesen 32 529 darab kiosztott eszközböngésző azonosító alapján a mintában az egyes eszközböngészők azonosításához $\Delta S = -\log_2 1/32\,529 = 13{,}81 \approx 14$ bit szükséges.

A Facebook azonosító esetében a $\Delta S = 0$, hiszen ha elérhető, akkor az adott eszközböngészőt használó felhasználót egyértelműen azonosítja. Tehát ha a ΔS-vel jelölt entrópia csökkentő tényező nullához közeli szám, akkor az jó bizonytalanság csökkentő erővel bír. A táblázatban szereplő nullánál nagyobb érték (0,28) azzal magyarázható, hogy néhány látogató több alkalommal is engedélyezte az adatgyűjtő alkalmazás számára az adataihoz való hozzáférést.

Amíg a rengeteg egyedi elemmel rendelkező telepített betűtípusok ΔS értéke nullához közeli, addig a mindössze 2 értéket felvevő jellemzők esetében (DNT, Flash

blokkoló jelenléte vagy a telepített PDF olvasó) értéke igen magas, hiszen ezen változó bizonytalanságcsökkentő ereje gyenge.

Ezt követően a paraméterek csoportjának bizonytalanság csökkentő erejét teszteltem. A kapott eredményeket az alábbi táblázatban foglaltam össze:

Paraméter	Lementett értékek száma	Egyedi elemek száma	Bizonytalanság csökkentő tényező (ΔS) (bit)
Böngésző környezetének jellemzői[22]	32 529	3 141	3,37
Szoftver és hardver környezet jellemzői[23]	32 529	10 560	1,62
Hardver környezet és a böngésző jellemzői[24]	32 529	14 594	1,15
Egy átlagos munkamenet ideje alatt állandó paraméterek[25]	32 529	17 845	0,86

16. táblázat: Lementett változók bizonytalanság csökkentő ereje

A táblázatból látható, ahogy egyre több releváns paramétert veszünk be, az egyedi elemek száma nő.

4.5. Összegzés

Az internet elképesztő fejlődésének köszönhetően nemcsak új iparágak jöttek létre, hanem a már létezők további fejlődésére is jótékony hatással van. Ezek egyik nyertese a marketing. Az online reklámok hatékonysága jóval magasabb, mint a hagyományos reklámoké, mivel az online hirdetésekkel pontosabban érhetőek el a célcsoportok.

Ha az online kereskedelemben részt vevő cégek többet tudnak a vásárlóik ízléséről, az igényeikhez jobban igazodó célzott reklámokat küldhetnek számukra. (Escobido & Gillian, 2013) Ezt az információt a nyílt internetről nem minden esetben lehetséges összegyűjteni, ha mégis, akkor az igen költséges. Az effajta személyes

[22] HTTP UserAgent, HTTP Accept, HTTP Accept Encoding, HTTP Accept Language, HTTP Accept Charset, DNT, böngésző típusa és verziója, földrajzi pozíció

[23] Képernyő méretei, operációs rendszer, érintőképernyő, kapcsolódás típusa, telepített betűkészletek

[24] HTTP UserAgent, HTTP Accept, HTTP Accept Encoding, HTTP Accept Language, HTTP Accept Charset, DNT, böngésző típusa és verziója, képernyő mérete, operációs rendszer típusa és verziója, földrajzi pozíció, érintőképernyő, kapcsolódás típusa, telepített betűkészletek

[25] Képernyő mérete, HTTP UserAgent, HTTP Accept, HTTP Accept Encoding, HTTP Accept Language, HTTP Accept Charset, DNT, böngésző típusa és verziója, operációs rendszer típusa, földrajzi pozíció, érintőképenyő, Flash, a Flash blokkolt-e, Quicktime, Java, PDF olvasó, Acrobat Reader verziószáma, telepített betűkészletek

jellemzők a látogatók az online világban nyújtott viselkedéséből következtethetőek ki. (Kosinksi, Stillwell, & Graepel, 2013)

Érdekes ellentétek figyelhetők meg az emberek online viselkedésével kapcsolatban. A felhasználók féltik a személyes adataikat és mindenképpen szeretnék elkerülni azok kompromittálódását, még akkor is, ha azt a felhőben tárolják. Ezzel ellentétes az a folyamat, hogy a népszerű közösségi hálózatokon az emberek gyakorlatilag ellenszolgáltatás nélkül kiteregetik akár a legbelsőbb titkaikat is. Felmerülhet a kérdés, hogy hogyan várhatja el valaki, hogy a felhőbe feltöltött adatai biztonságban legyenek, ha az alapvető biztonsági ajánlásokat sem tartja be és esetleg más fórumokon tálcán kínálja a bűnözők számára az adataihoz való hozzáférés lehetőségét. A webes böngészéssel rengeteg adat elérhető a böngésző személyéről, az általa használt eszközökről szokásairól más egyéb személyes tulajdonságairól, amelyek alapján a látogatók profilozhatóvá válnak.

A feltáró jellegű kutatásom célja annak bemutatása, hogy az egy domain alól elérhető weboldalak számára a látogatók által használt böngészők és eszközökről elérhető adatok használhatóak-e a látogató azonosítására és követésére. A kutatásomban elemeztem, hogy a látogató eszközböngészőjéből hozzáférhető adatok milyen mértékben járulnak hozzá ehhez.

Egy oldalletöltés alkalmával akár néhány száz kilobyte-nyi adat hozzáférhető a meglátogatott webhely számára, amelyből adatbányászati módszerekkel értékes – akár személyhez köthető – információk nyerhetőek ki, valamint ezen adatok segítségével lehetséges az egyének beazonosítása és követése, személyes adataik felhasználása nélkül. A rendelkezésre álló paramétereket a következő négy csoportba rendeztem: a böngésző környezetének jellemzői, a szoftver és hardver környezet jellemzői, a hardver környezet és a böngésző jellemzőit és az egy átlagos munkamenet ideje alatt állandó paramétereket. A kutatás eredményeképpen azt tapasztaltam, hogy minél több változót használok a felhasználó azonosításához, annál nagyobb vizsgálatba bevont változók együttes bizonytalanság eloszlató képessége. Ez azt jelenti, hogy egy ismert sokaság elemei közül nagy bizonyossággal be tudjuk azonosítani az egyén által használt eszközböngészőt, feltéve, hogy a munkamenet ideje alatt állandó paraméterei ismertek. Ebből az is következik, hogy amennyiben meg szeretnénk nehezíteni a követőink dolgát, célszerű proxy-n keresztül minél kevesebb paramétert megosztani a követő weboldalak és harmadik felek számára.

5. AZ EGYETEMI POLGÁROK PSZICHOLÓGIAI JELLEMZŐI ADATVÉDELMI SZEMPONTBÓL

Az Egyetemi polgárok pszichológiai jellemzőinek vizsgálata során a Budapesti Corvinus Egyetem polgárainak a legnagyobb közösségi oldal (Facebook) által összegyűjtött, kinyilvánított preferenciákból kinyert személyes tulajdonságait elemzem. A kutatás keretében megismerhető, hogy a felhasználók által önként szolgáltatott adatokból miként lehetséges személyes jellemzőket kinyerni, megismerhető az Egyetemi polgárok személyes adatainak megosztási hajlandósága, valamint bemutatja a hozzáférhető személyes adatok sokaságát. Az egyéneket a személyes tulajdonságaik alapján klaszterekbe rendeztem, majd az eredményeimet összehasonlítottam a myPersonality Project kutatásának adatgyűjtési fázisának eredményeképpen létrejött mintából képzett klaszterekkel.

A University of Cambridge Psychometrics Center kutatói az online profilozás mérföldkövét tették le 2012-ben, amikor 58 000 olyan önkéntes Facebook felhasználóval pszichológiai teszteket töltettek ki, majd ezt követően elemezték a látogatók Facebook-on található személyes adataikat is. Az elemzések eredményeinek egybevetését követően fejlesztettek egy mindenki számára elérhető API-t, amely a Facebook-os like-ok alapján képes meghatározni a vizsgálat alanyának pszichológiai jellemzőit. (Kosinksi, Stillwell, & Graepel, 2013)

2015. januárjában publikálták egy újabb kutatásuk eredményét, amely szerint a számítógépes személyiségelemzésnél jóval pontosabb eredményt képes elérni, mint a hús-vér emberek, mivel a számítógép számítási képessége mára jelentősen felülmúlja az emberekét, valamint mentes az előítéletektől és kevesebbet is hibázik. A kutatás során 86 000 emberrel töltettek ki egy 100 kérdésből álló személyiségtesztet, amely alapján képesek voltak lemérni a résztvevők 5 fő személyiségi jegyét (Big 5). A Big 5 személyiségi jellemzők leírása az 5.2 A Big 5 – személyiségi jellemzők fejezetben olvasható. A kérdőívet kitöltők eredményét összevetették a Facebook-os aktivitásukkal és ennek eredményeképpen pontosabban tudták megjósolni az egyének pszichikai egészségét, politikai hovatartozást és tulajdonképpen az életük „eredményét”. Néhány esetben a módszer pontosabban írta le az egyént, mint ő saját maga.

Az eredmények felhasználásával fejlesztettek egy alkalmazást a Facebook-ra, amelynek a „képességei” megdöbbentőek. A kutatócsapat által kitalált modell számára inputként csak 10 Facebook-os like-ra van szüksége, hogy pontosabban legyen képes felmérni az elemzendő személyiségi jegyeit, mint egy kolléga, 70 like-ra, hogy pontosabban, mint egy szobatárs, 150 like-ra, hogy jobban, mint egy családtag és pusztán 300 like-ra, hogy pontosabban egy házastársnál. (Youyou, Kosinki, & Stillwell, 2015)

5.1. A myPersonality Project

A University of Cambridge kutatói a 2012-es kutatásuk adatgyűjtési szakaszában lementett egyének jellemzőinek anonimizált adatait kutatók számára hozzáférhetővé tették. Így lehetségessé vált a saját kutatásom keretében gyűjtött adatokkal való összevetése.

A számítógépek megjelenésével lehetővé vált a nagy mennyiségű természetes nyelvi szövegek elemzése, amelyekből következtetni lehet az előbb említett alapvető emberi jellemvonásokra, a különböző helyzetekben valószínű reakcióra és viselkedésmintákra. (Poria, Gelbukh, Agarwal, Cambria, & Howard, 2013) Szintén bizonyított a kapcsolat a különböző személyiségi jellemzőkkel bíró egyének online platformokon (Facebook) való megjelenése között. (Golbeck, Robles, & Turner, Predicting personality with social media, 2011)

A myPersonality Project kutatást követően sok olyan publikáció született, amelyek a kutatás eredményeképpen megszületett Facebook Like – pszichológiai jellemző adatbázison alapult. A kutatások egy része valamilyen feltételezést igazolt:

- Marketing kutatások igazolták, hogy azon egyének, akik a magas nyitottság személyiségjegy értékkel rendelkeznek, innovátorok és sok esetben képesek másokat befolyásolni. (Kosinski, és mtsai., 2012)
- A Facebook’s Gross National Happiness Index (Facebook Össznemzeti Boldogság Index - FGNH) egy számba sűríti a Facebook-os üzenetekben lévő szavak pozitív vagy negatív töltetét. Kiderült, hogy egy a szám a nemzeti ünnepek környékén tetőzik és nyomasztó események vagy megemlékezések alkalmával csökken. (Stillwell D. , Kosinski, Rust, & Wang, 2012)

A Projekt egy másik része pedig az online profilépítés témakörét boncolgatja:

- Golbeck Jennifer, Robles Cristina és Turner Karen „Predicting personality with social media” című cikkükben közöltek egy olyan metódust, amely képes megjósolni az egyén személyiségi jellemzőit a Facebook-on megjelenő publikus profilja alapján. A kutatások azt bizonyítják, hogy van kapcsolat az ember személyiségi jellemzői és online viselkedése között. (Golbeck, Robles, & Turner, Predicting personality with social media, 2011)
- A Michal Kosinski és társai által írt „Facebook and Privacy: The Balancing Act of Personality, Gender and Relationship Currency” cikk kategorizálja a felhasználókat a személyiségi jellemzőik alapján, amelyből kiderül, hogy az adatvédelem tudatos felhasználók, annál magasabb a nyitottság és az extrovertáltság személyiségi jellem értéke. A nemek közt nincs különbség a megosztás tekintetében, ámbár a nők hajlamosabbak az elővigyázatosságra és kevesebb információt osztanak meg magukról. (Kosinski, és mtsai., 2012)

5.2. A Big 5 – személyiségi jellemzők

A „Big 5” (magyarul „Nagy Ötös”) személyiségi jellemzők az akadémiai pszichológia egyik elfogadott és széles körben használt módszere az általános emberi jellemzők leírására. Az 5 alapvető emberi személyiségjegyet - amelynek létezését számos lélektani és biológiai háttérkutatás is igazolta - faktoranalízissel, empirikus mérési adatok elemzésével és a természetes nyelvből szisztematikusan összegyűjtött és rendszerezett tulajdonságok elemzése útján nyerték ki. (Mirnics, 2006)

Az előző fejezetben ismertetett myPersonality Project eredményeképpen létrehozott pszichológiai API a vizsgált személyek személyiségjegyeit a Big 5 metodológia szerint adja vissza, emiatt a kutatásaim során én is ezt használtam.

A Big 5 által használt alap személyiségi jellemzők rendre:

- **Nyitottság** (openness) személyiségjegy az új kalandokra, ötletekre és esztétika befogadására való hajlandóság, kíváncsiság és képzelőerő.
- **Lelkiismeretesség** (conscientiousness) a rendszerezettség kedvelése. A lelkiismeretes emberek rendszerint ambiciózusak, talpraesettek és kitartóak.
- **Extrovertált** (extraversion) személyiségjeggyel rendelkező egyének társaságkedvelők, barátságosak és szociálisan aktívak. Rendszerint lendületesek, beszédesek, imádnak a figyelem középpontjában lenni és könnyen szereznek új barátokat. Az introvertált személyekre ezen

tulajdonságok ellentettjei jellemzőek: jobban kedvelik a saját maguk társaságát és keresik a csendes, ingerszegény környezetet.

- A **barátságos** (agreeableness) emberek bizakodók, önzetlenek és gyengédek. Azon egyének, akik erre a kategóriára magas értéket kapnak, nagy valószínűséggel barátságosak, együttérzők és tapintatosak.
- A **neurotikusság** (neuroticism) jellemző az érzelmi felelősséggel bíró és impulzív személyekre. Esetenként gyors hangulatváltozások, negatív érzelmek (bűntudat, düh, szorongás, depresszió) jellemzik. (Kosinski, Bachrach, Kohli, Stillwell, & Graepel, 2013)

5.3. Apply Magic Sauce

A myPersonality Project végeredményeképpen létrejött Facebook Like – személyiségjegy adatbázis köré épített API neve: „Apply Magic Sauce". A RESTful API-n keresztül elérhető szolgáltatás Facebook Like-okból képes megállapítani az egyén pszichológiai-demográfiai jellemzőit és egy regisztráció ellenében bárki számára elérhető.

„Apply Magic Sauce translates individuals' digital footprints into detailed psychological profiles"

17. ábra: Apply Magic Sauce - Prediction API alcíme

(The Psychometrics Centre, 2013)

Az API inputként a vizsgálandó Facebook felhasználó Like azonosítóinak halmazát várja és válaszként az alábbi személyes adatokat adja vissza:

Adat	Pearson-féle korrelációs együttható[26]
BIG5	0,35 – 0,5
Élettel való elégedettség	0,17/0,44
Intelligencia	0,47
Kor	0,75
Nem	0,93
Szexuális orientáció	férfiak esetében: 0,88 nők esetében: 0,75
Érdeklődési kör	0,72
Politikai beállítottság	0,79
Hitvallás	0,76
Családi állapot	0,67

6. táblázat: Apply Magic Sauce API-ja által visszaadott személyes adatok

(The Psychometrics Centre, 2013)

Innen már csak egy lépés az, hogy a webboltokba bejelentkező vásárló személyiségét feltérképezzék pszichológiai API segítségével, majd a vásárló személyiségjegyei alapján ajánlanak termékeket számára vagy a híroldalakon megjelenő híreket a detektált érdeklődési körének megfelelően jelenítik meg.

Alessandro Acquisti (Carnegie Mellon University) 2013. júniusában tartott TED-es előadásában az előző gondolat valóságtartalmát igazolta a kutatásai során. A megnövekedett online aktivitás és közösségi hálózatra való információfeltöltés következtében bármilyen személyes információ, érzékeny információvá válhat. (Alessandro, 2013)

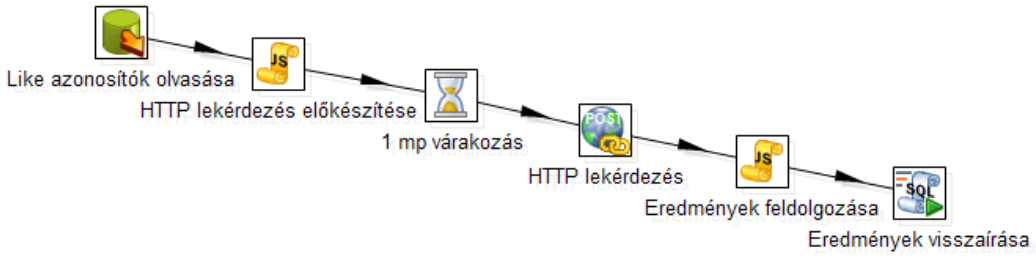
5.4. Az adatelemzés

Az online környezetben a valós személyek a személyiségi jellemzőiknek megfelelően viselkednek. (Kosinksi, Stillwell, & Graepel, 2013) Kutatásom keretében

[26] A Pearson féle korrelációs együtthatóval jellemezhető a változók közti lineáris kapcsolat szorosságát. (0,5 felett erősnek, 0,3-0,5 között közepes erősségű, és 0,1-0,3 között pedig gyenge erősségű) (Hunyadi & Vita, 2006)

azt szemléltetem, hogy a közösségi hálózatokon tanúsított viselkedések lenyomatáról elérhető információkból személyiségelemzést lehet elvégezni.

Készítettem egy Pentaho alkalmazást, amely kiolvassa a látogató like-jait az adatbázisból, elküldi azt az Apply Magic Sauce API-nak, majd a visszakapott eredményeket szintén elmenti a látogatóhoz társítva.



18. ábra: Facebook Like-okat személyiségi jegyekké feloldó Pentaho alkalmazás (saját szerkesztés)

Az API csak bizonyos számú Like esetén képes az illető személyiségi jellegzetességeit megjósolni, ezt a feltételt 48 egyén teljesítette a 139-ből. Ezt követően kezdtem neki az adatok elemzésének. Klaszterképző eljárásokkal hasonló egyéneket kerestem az így kapott mintában, majd az eredményt összevetettem a myPersonality Project adatgyűjtő fázisában lementett adatokból képzett klaszterekkel.

5.4.1. A myPersonality Project adatai

A myPersonality Project adatbázisából a Big 5, a demográfiai, az IQ, az élettel való megelégedettség, a vallás és a politikai nézetek adattáblák adatait töltöttem le elemzési célból. A hozzá tartozó leírás szerint a letöltött mintában szereplő egyének korának átlaga 23,55 év. A mintában szereplő egyének 79,06%-a USA, Egyesült Királyság, Kanada és Ausztrália területéről vett részt a kísérletben. (Kosinksi, Stillwell, & Graepel, 2013)

5.4.2. A statisztikai elemzés

Először leíró statisztikákkal jellemeztem az elemzendő mintákat. Az eredmények táblázata az általam gyűjtött és myPersonality Project keretében gyűjtött átlagát és 95%-os konfidencia intervallumát tartalmazza. (10-13. táblázat)

Big 5	myPersonality adat			saját gyűjtés		
	Átlag	95% konfidencia intervallum		Átlag	95% konfidencia intervallum	
Barátságosság	,6281	,6207	,6354	,3413	,3030	,3795
Lelkiismeretesség	,5756	,5676	,5835	,4728	,4235	,5220
Nyitottság	,7736	,7673	,7800	,4509	,4196	,4822
Neurotikusság	,4473	,4377	,4569	,3245	,2892	,3597
Extrovertáltság	,5323	,5227	,5418	,3765	,3431	,4098

7. táblázat: A Big 5 jellemzőinek átlaga és 95%-os konfidencia intervalluma

Családi viszony	myPersonality adat			saját gyűjtés		
	Átlag	95% konfidencia intervallum		Átlag	95% konfidencia intervallum	
Egyedülálló	,4651	,4429	,4873	,5480	,5355	,5605
Házas	,0857	,0733	,0982	,1449	,1349	,1550
Kapcsolatban	,1828	,1656	,1999	,3071	,2980	,3161

8. táblázat: A Big 5 jellemzőinek átlaga és 95%-os konfidencia intervalluma

Politikai nézet	myPersonality adat			saját gyűjtés		
	Átlag	95% konfidencia intervallum		Átlag	95% konfidencia intervallum	
Konzervatív	,0103	,0058	,0147	,3171	,2917	,3425
Liberális	,0277	,0204	,0350	,4072	,3783	,4361
Nem érdekli a politika	,0154	,0099	,0209	,1676	,1538	,1814
Libertariánus	,0062	,0027	,0096	,1081	,0979	,1183

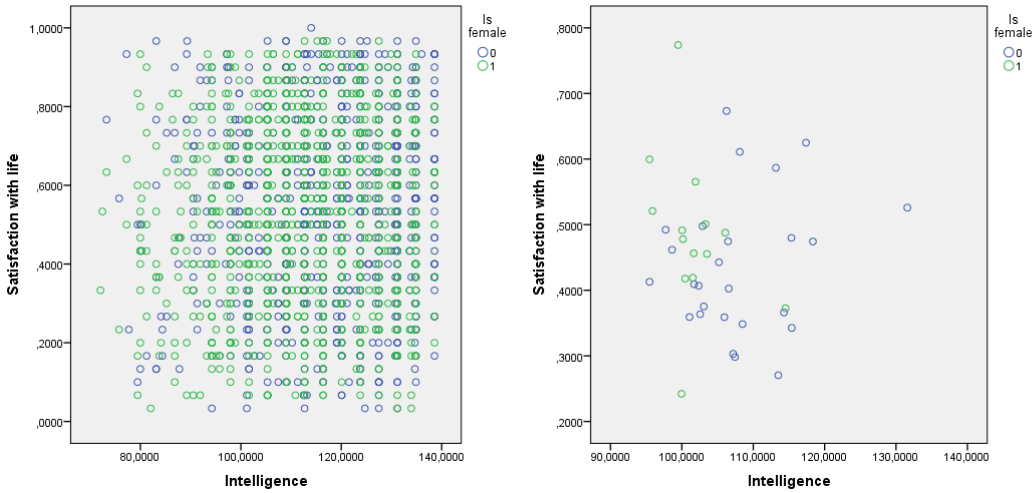
9. táblázat: A Big 5 jellemzőinek átlaga és 95%-os konfidencia intervalluma

Egyéb változók	myPersonality adat			saját gyűjtés		
	Átlag	95% konfidencia intervallum		Átlag	95% konfidencia intervallum	
Kor	28,09	27,61	28,58	26,11	25,67	26,54
Barátok száma	106,10	97,63	114,57	594,50	513,35	675,65
Intelligencia	114,77	114,12	115,4	104,99	103,09	106,97
Élettel való elégedettség	,5337	,5231	,5442	,4536	,4180	,4891

10. táblázat: A Big 5 jellemzőinek átlaga és 95%-os konfidencia intervalluma

A leíró statisztikák táblázatából leolvasható, hogy a konfidencia intervallumoknak egy esetben sincs metszetük és a mintákban szereplő értékek átlagai szignifikánsan különböznek.

A myPersonality mintájában található egyének korának átlaga több, mint a saját gyűjtésem egyéneinek kora, ami az Egyetemre járó diákok korával magyarázható. Érdekes, hogy a saját gyűjtésemben található egyének intelligenciája és az élettel való megelégedettsége alacsonyabb, azonban jelentősen több ismerősük van a közösségi hálón. Az összes Big 5 érték átlaga alacsonyabb a saját gyűjtésem mintájában, mint a myPersonlity Project adatai esetében. Szintén meglepő, hogy a magyar mintában szereplő egyének politikai nézeteinek értékei jelentősen magasabbak. A politikától való tartózkodás is egyértelmű a magyar adatok esetében.



19. ábra: Az intelligencia és az élettel való elégedettségét megjelenítő scatterplot diagram

Az intelligencia és az élettel való elégedettség változók kapcsolatának vizsgálatához az összetartozó értékeket scatterplot diagramon ábrázoltam. A diagram a két változó függetlenségét magyarázza, még akkor is, ha nemek szerint különválasztjuk az egyéneket.

5.4.3. Klaszterelemzés

A lementett személyiségi jegyek és demográfiai adatok halmazát klaszterező eljárásokkal elemeztem, amelynek célja a megfigyelési egységek homogén csoportokba rendezése a kiválasztott változók alapján. (Cser & Fajszi, 2004) Az elemzést a myPersonality Project adatgyűjtési fázisában összegyűjtött adatokkal kezdtem, majd azt a saját adatgyűjtésből származó adatokra is lefuttattam. K-közép és hierarchikus klaszterképző eljárások eredményének összevetéséből arra a megállapításra jutottam, hogy a kapott klaszterek stabilak.

K-közép klaszterképzés

A K-közép klaszterképző eljárás egyik hátránya, hogy eleve feltételezi a klaszterek létezését a mintában. Kezdetben 2, majd 3 és 4 klasztert feltételezve alakítottam ki a csoportokat. A távolság mérésére euklideszi távolság módszerét választottam, a kezdő középpontokat pedig az SPSS adta meg.

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Big 5 Barátságosság	8,910	1	,020	1946	434,660	,000
Big 5 Lelkiismeretesség	8,901	1	,024	1946	366,586	,000
Big 5 Nyitottság	1,892	1	,018	1946	106,342	,000
Big 5 Neurotikusság	43,019	1	,022	1946	1945,223	,000
Big 5 Extrovertáltság	30,532	1	,026	1946	1158,872	,000

11. táblázat: A myPersonality Big 5 jellemzőinek ANOVA táblája

A myPersonality Project Big 5 adatain elvégzett klaszterképzés eredménye a 14. táblázatban látható. A létrehozott 2 klaszter értékei szignifikánsan eltérnek egymástól, amit az alacsony p-érték és az F-tesztek is alátámasztanak.

Ezt követően az egyéneket a politikai nézeteik alapján klaszterezni. A kapott eredmények a 15. táblázatban láthatóak. Az eredmények alapján – a liberális nézeteket valló egyéneken kívül - a klaszterek nem szignifikánsak, mivel a p-érték és az F-tesztek értékei az 5%-os szignifikancia szint felett vannak.

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Konzervatív	,006	1	,010	1946	,576	,448
Liberális	52,503	1	,000	1946	.	.
Nem érdeklődik	,013	1	,015	1946	,868	,352
Libertariánus	,002	1	,006	1946	,344	,558

12. táblázat: A myPersonality politikai jellemzőinek ANOVA táblája

A 16. táblázat alapján a családi állapot jellemzők alapján megfelelő minőségben lehet klasztereket képezni.

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Egyedülálló	39,511	1	,229	1946	172,739	,000
Házas	152,683	1	,000	1946	.	.
Kapcsolatban	6,100	1	,146	1946	41,678	,000

13. táblázat: A myPersonality családi állapot jellemzőinek ANOVA táblája

A klaszterek képzése során elmentettem a klaszter tagságot, ami lehetővé teszi azok összevetését. Az összevetést kereszttábla (crosstab) módszerrel végeztem, az eredményt a 17. táblázat tartalmazza. A B_clusters a Big 5 értékei alapján készített klasztereket veti össze a S_cluster a családi állapot alapján készített klasztertagsággal.

B_clusters * S_clusters Crosstabulation

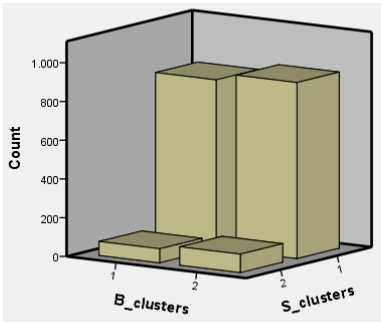
Count

		S_clusters		
		1	2	Total
B_clusters	1	872	72	944
	2	909	95	1004
Total		1781	167	1948

14. táblázat: A klaszter tagság kereszttáblája

A kapott eredmény alapján kijelenthető, hogy a klaszterek függetlenek egymástól, amelyet ellenőrizhetünk a Pearson Chi-Square (2,090) értékével és a hozzá tartozó p-értékkel (0,148). Mivel a kapott p-érték nem kisebb az 5%-os szignifikancia

szintnél, a nullhipotézis nem vethető el, amely a 3 dimenziós hisztogramról is jól látszik.



20. ábra: A klaszterek 3 dimenziós hisztogramja

A politikai nézet jellemzők nem megfelelő klaszterképző ereje miatt azokat nem vettem bele a végső klaszterképző algoritmusba, azonban a Big 5 és a családi állapot jellemzőket azok függetlensége miatt igen. Az algoritmust lefuttattam 2, 3 és 4 klaszter esetére, amiből kiderült, hogy a 4 klaszter a megfelelő választás, mivel ebben az esetben lesz a változók szétválasztó ereje a legnagyobb.

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Big 5 Barátságosság	,065	3	,025	1944	2,612	,050
Big 5 Lelkiismeretesség	,236	3	,029	1944	8,265	,000
Big 5 Nyitottság	,106	3	,019	1944	5,683	,001
Big 5 Neurotikusság	,070	3	,044	1944	1,586	,191
Big 5 Extrovertáltság	,151	3	,042	1944	3,610	,013
Egyedülálló	161,542	3	,000	1944	.	.
Házas	50,894	3	,000	1944	.	.
Kapcsolatban	96,980	3	,000	1944	.	.

15. táblázat: A Big 5 és a családi állapot klasztereinek ANOVA táblája

A családi állapot jellemzőknek nincs F-értéke és p-értéke, mert azok alapján az egyének teljes mértékben elkülöníthetők. A végső klaszter középpontok a 19. táblázatban tekinthetőek meg.

Final Cluster Centers

	Cluster			
	1	2	3	4
Big 5 Barátságosság	,6126	,6346	,6259	,6418
Big 5 Lelkiismeretesség	,5712	,5715	,5566	,6334
Big 5 Nyitottság	,7811	,7678	,7888	,7410
Big 5 Neurotikusság	,4566	,4385	,4639	,4453
Big 5 Extrovertáltság	,5250	,5218	,5583	,5550
Egyedülálló	0	1	0	0
Házas	0	0	0	1
Kapcsolatban	0	0	1	0

16. táblázat: Végső klaszter középpontok

Hierarchikus klaszterképzés

A kapott eredmények validálása miatt a hierarchikus klaszterképző eljárást is lefuttattam a kiválasztott jellemzőire. A klaszterképző eljárás paramétereiben Ward metódusát, a távolság számítására euklideszi távolságot használtam. A Ward-féle varianciamódszer azokat a klasztereket vonja össze, melyek esetében az összevonás során a belső szórásnégyzet növekedése a legkisebb lesz és ezzel egyenlő elemszámú klaszterek kialakítására törekszik. (Kovács E. , 2014) A hierarchikus klaszterezés előnye, hogy nem szükséges előre tudnom a klaszterek számát, mert az leolvasható az algoritmus eredményeképpen létrejött dendrogramról. Esetemben a hierarchikus klaszterezés igazolta a megfelelően kiválasztott 4 klasztert.

A két különféle klaszterképző eljárással klaszterekbe sorolt egyének tagságát elmentettem, majd egy kereszttábla elemzéssel vetettem össze őket.

Cluster Number of Case * Ward Method Crosstabulation

Count

		Ward Method				
		1	2	3	4	Total
Cluster Number of Case	1	0	519	0	0	519
	2	906	0	0	0	906
	3	0	0	0	356	356
	4	0	0	167	0	167
Total		906	519	167	356	1948

17. táblázat: Klaszterek kereszttáblája

A kereszttábla elemzés eredménye a 20. táblázatban tekinthető meg, amely igazolta, hogy a két klaszterképző eljárással képzett klaszterek megegyeznek, tehát azok stabilak.

Saját adatok elemzése

A klaszterképző eljárásokat az általam összegyűjtött adatokra is lefuttattam. A futtatások alkalmával az algoritmusokat az előzőekben ismertetett paraméterekkel használtam és a Big 5, politikai nézet és a családi állapot változókat vettem be a vizsgálatba. Az eredményeket a 18. *táblázat: Big 5 klaszterek ANOVA táblája,* a 19. *táblázat: Politikai nézet ANOVA táblája* és a 20. táblázat*: Családi állapot ANOVA táblája* tartalmazza.

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Big 5 Barátságosság	,067	1	,013	38	5,140	,029
Big 5 Lelkiismeretesség	,611	1	,008	38	74,019	,000
Big 5 Nyitottság	,038	1	,009	38	4,317	,045
Big 5 Neurotikusság	,069	1	,011	38	6,454	,015
Big 5 Extrovertáltság	,064	1	,009	38	6,747	,013

18. táblázat: Big 5 klaszterek ANOVA táblája

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Konzervatív	,122	1	,003	38	37,399	,000
Liberális	,226	1	,002	38	93,156	,000
Nem politizál	,002	1	,002	38	1,074	,307
Libertariánus	,007	1	,001	38	7,680	,009

19. táblázat: Politikai nézet ANOVA táblája

A 14. táblázatban a „Nem politizál” változó kivételével mindegyik változó szignifikáns és hozzájárul a klaszter kialakításához.

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Egyedülálló	,032	1	,001	38	43,480	,000
Házas	,013	1	,001	38	18,541	,000
Kapcsolatban	,004	1	,001	38	6,064	,018

20. táblázat: Családi állapot ANOVA táblája

A Big 5 és a családi állapot klaszterek összehasonlításának eredménye a 21. táblázatban látható.

B_clusters * S_clusters Crosstabulation

Count

		S_clusters		
		1	2	Total
B_clusters	1	5	5	10
	2	9	21	30
Total		14	26	40

21. táblázat: Családi állapot kereszttábla

A táblázat a változók függetlenségét sugallja, de a mintában lévő elemek alacsony száma miatt ebben nem lehetünk biztosak. A függetlenségi vizsgálat eredménye 1,319 Pearson Chi-Square érték és 0,251 p-érték, így a függetlenség hipotézisét nem vethetjük el 5%-os szignifikancia szinten.

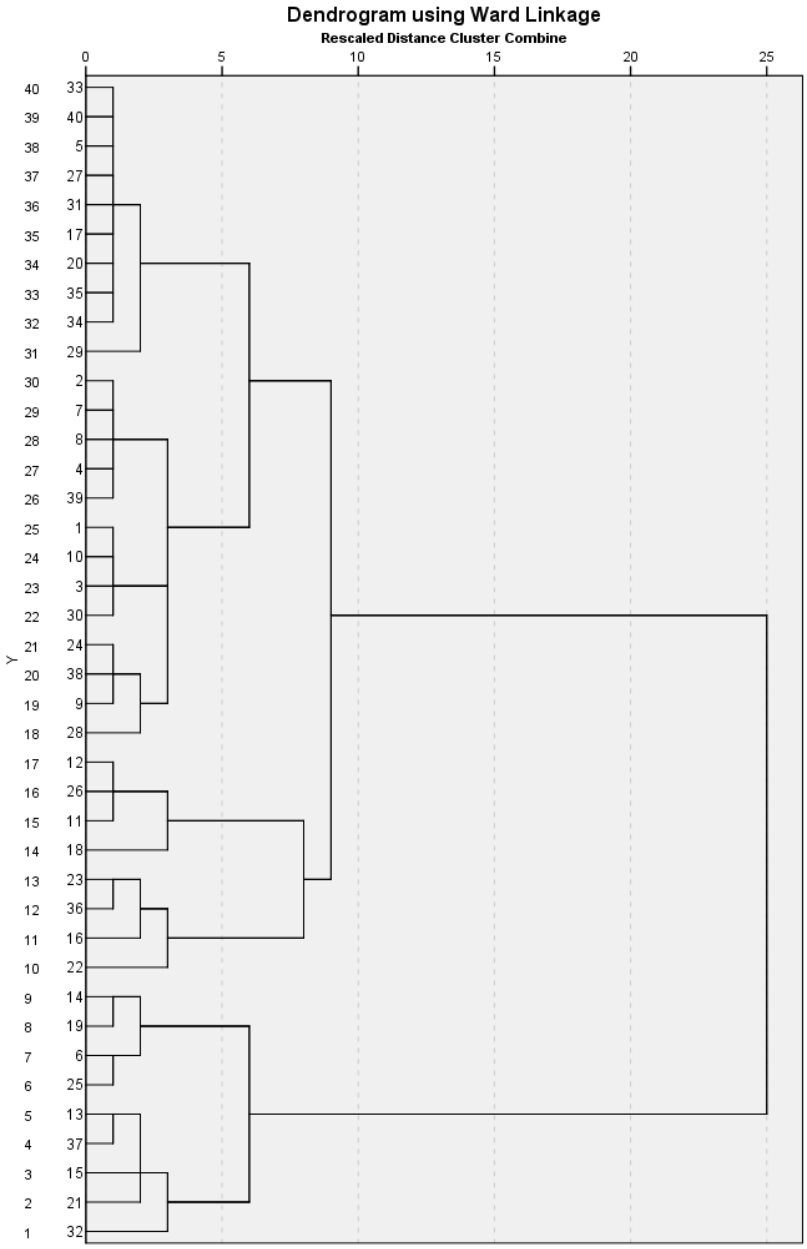
Ezt követően a K-közép klaszterképző algoritmus lefuttattam 2, 3 és 4 klaszterre is. Az eredmények szerint 3 klaszter esetében az összes elemzésbe bevett változó szignifikáns.

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Big 5 Barátságosság	,145	2	,007	37	20,077	,000
Big 5 Lelkiismeretesség	,343	2	,006	37	52,967	,000
Big 5 Nyitottság	,066	2	,007	37	10,097	,000
Big 5 Neurotikusság	,019	2	,012	37	1,631	,210
Big 5 Extrovertáltság	,020	2	,010	37	1,949	,157
Egyedülálló	,004	2	,001	37	3,096	,057
Házas	,003	2	,001	37	3,630	,036
Kapcsolatban	,004	2	,001	37	7,272	,002

22. táblázat: A BIG5 és a családi állapot változóinak ANOVA táblája

A hierarchikus klaszterelemzésnél használatos dendrogramról leolvasható, hogy 2 vagy 3 klaszter képzése a megfelelő választás. Az összehasonlítás során 2 klaszter képzése esetén az ANOVA táblából leolvasható, hogy a házas (Married) változó nem szignifikáns, ezért 3 klasztert képeztem.



21. ábra: Hierarchikus klaszterelemzés dendrogramja

A K-közép és a hierarchikus klaszterezés eredményeinek összehasonlításának végeredménye a 23. táblázatban látható: a két klaszterképző eljárás majdnem ugyanazt az eredményt adta, így kijelenthető, hogy a képzett klaszterek stabilak.

Ward Method * Cluster Number of Case Crosstabulation

Count

		Cluster Number of Case			
		1	2	3	Total
Ward Method	1	0	22	1	23
	2	8	1	0	9
	3	0	1	7	8
Total		8	24	8	40

23. táblázat: Klaszter tagság ellenőrzés

5.4.4. Az eredmények kiértékelése

A végső klaszter középpontokat tartalmazó táblázatokat hasonlítottam össze, amelyekből megállapítható, hogy a Corvinus Egyetem polgárai milyen jellegzetességekkel rendelkeznek.

Final Cluster Centers

	Cluster			
	1	2	3	4
Big 5 Barátságosság	,6126	,6346	,6259	,6418
Big 5 Lelkiismeretesség	,5712	,5715	,5566	,6334
Big 5 Nyitottság	,7811	,7678	,7888	,7410
Big 5 Neurotikusság	,4566	,4385	,4639	,4453
Big 5 Extrovertáltság	,5250	,5218	,5583	,5550
Egyedülálló	0	1	0	0
Házas	0	0	0	1
Kapcsolatban	0	0	1	0

24. táblázat: myPersonality Project adatainak K-középpontú klaszterei

Final Cluster Centers

	Cluster		
	1	2	3
Big 5 Barátságosság	,2284	,4107	,2457
Big 5 Lelkiismeretesség	,7207	,4379	,3293
Big 5 Nyitottság	,4235	,4217	,5659
Big 5 Neurotikusság	,2667	,3317	,3605
Big 5 Extrovertáltság	,4400	,3597	,3632
Egyedülálló	,5749	,5376	,5522
Házas	,1442	,1532	,1207
Kapcsolatban	,2809	,3091	,3271

25. táblázat: a saját adatok K-középpontú klaszterei

A myPersonality Project adatokból képzett klaszterek értelmezése:

- **Valódi életet élők:** Ámbár nyitottnak tűnnek, nem sokat fednek fel magukból a Facebook-on, valószínűleg kevés időt töltenek a közösségi hálózaton.
- **Magányosak:** Nincsenek kapcsolatban és nem élnek társasági életet a Facebook-on.
- **Beszédesek:** Ez a kapcsolatban lévő emberek rendelkeznek a legmagasabb nyitottság, neurotikusság és extrovertáltság értékével. Leginkább ők érzik szükségét a közösségi hálózaton történő megosztásnak.
- **Komolyak:** Házas, kompromisszumképes és lelkiismeretes csoport.

A Corvinus Egyetem polgárainak adatait elemezve az alábbi 3 csoportot állapítottam meg:

A táblázatból leolvasható az alábbi 3 klaszter és a feltételezhető magyarázatuk:

- **Elmélkedők**: Nagy valószínűséggel nincsenek kapcsolatban, de meglehetősen extrovertáltak és lelkiismeretesek.
- **Komolyak**: A legkellemesebb személyiséggel rendelkezők, szintén lelkiismeretesek, valószínűleg házasok vagy komoly kapcsolatban vannak.
- **Beszédesek**: Nagyon nyitott és extrovertált személyek, akik valószínűleg kapcsolatban vannak.

5.5. Összegzés

A myPersonality Project keretében összegyűjtött nemzetközi adatokban és a saját adatgyűjtés eredményeképpen létrejött adatokban klasztereket képeztem nem felügyelt tanulási algoritmusokkal, majd a kapott eredményeket összevetettem. A kutatás eredményeképpen megállapítható, hogy a különböző forrású adatokban hasonló egyének csoportjai vettek részt, annak ellenére, hogy a myPersonality Project keretében gyűjtött adatok főként az angolszász országok polgárai, a saját kutatásomban pedig a Budapesti Corvinus Egyetem polgárai vettek részt.

A kutatásom idejében a Facebook like-ok pszichológiai elemzése során kinyerhető az egyénre jellemző Big5 személyiségi jellemzők, az élettel való elégedettség, intelligencia, kor, nem, szexuális orientáció, érdeklődési kör, politikai beállítottság, hitvallás és családi állapot. A pszichológiai API a Facebook like-okból visszaadott változók bizonyosságát a Pearson-féle korrelációs együtthatóval jellemzi,

amely szerint többségük (kor, nem, szexuális orientáció, érdeklődési kör, politikai beállítottság, hitvallás és családi állapot) igen erős[27] lineáris kapcsolatot, míg néhány (Big5, élettel való elégedettség és intelligencia) paraméter közepesen erős kapcsolatot mutat. A különböző forrású mintákon K-közép és hierarchikus klaszterképző eljárásokat futtattam, amelynek eredményeképpen létrejött klaszterek mindkét módszer esetében ugyanazokat eredményezte, így azok stabilnak mondhatók.

A vizsgált mintákban (myPersonality Project és a Corvinus Egyetem polgárai) kettő nagyon hasonló klaszter lelhető fel: a magas nyitottság, neurotikusság és extrovertáltság értékkel rendelkező tartalommegosztók, akik nagy valószínűséggel nincsenek kapcsolatban és a komoly kapcsolatban lévő, kompromisszumra képes, lelkiismeretesek egyének. Emellett a myPersonality Project mintájából kimutatható még a közösségi hálózaton kevés időt töltő egyének csoportja, akik feltételezhetően a Corvinus Egyetem polgárainak mintájában is jelen vannak.

[27] A Pearson-féle korrelációs együttható értéke magasabb, mint 0,67

6. SZEMÉLYES INFORMÁCIÓ KINYERÉSE WEBES ADATOKBÓL

A kutatók egyetértenek abban, hogy az egyén online tanúsított viselkedéséből és kinyilvánított preferenciáiból a személyes jellemzői kikövetkeztethetők. (Kosinski, Stillwell, & Graepel, 2013)

Az *Az Egyetemi polgárok pszichológiai jellemzői adatvédelmi* szempontból című kutatásom során a Facebook-os adatokból kinyert személyes jellemzőket a látogatók adataihoz társítva elmentettem egy adatbázisba. Ezt követően az online környezetben tanúsított viselkedés és a környezeti jellemzőkből következtettem a látogatók személyes tulajdonságaira. A kutatásban vizsgált mintában a személyhez nem köthető változók az attribútum halmaz elemei, a személyes tulajdonságok pedig a potenciális osztályváltozók.

A kutatásom során több klasszifikációs algoritmust, többek között a népszerű Apriori algoritmust is használtam, amely nagy adathalmazokban képes hatékonyan asszociációs szabályok keresésére. (Agrawal & Srikant, 1994) Az Apriori algoritmust több adatbányászati alkalmazásba is implementálták, én a Java alapú Wekát használtam. (Witten, Frank, & Hall, 2011)

Az Apriori algoritmus gyakori elemhalmazok elemzésével nyeri ki az asszociációs szabályokat. Az algoritmus kihasználja a gyakori elemhalmazok részhalmaza is gyakori axiómát, ezáltal az algoritmus hatékonyan tudja csökkenteni a kinyert szabályok számát, melyek közül csak azok lesznek érdekesek, amelyek egy minimális support (támogatottság) értéknél nem kevesebb. (Agrawal & Srikant, 1994) Már egészen kis minták esetében is rengeteg különböző asszociációs szabály nyerhető ki, az algoritmus csak azokat tartja meg, amely viszonylag nagy számú példányra érvényes. Ezen szabályokat érdekesnek hívjuk. Az algoritmus kimenetének minőségét az alábbi mutatókkal jellemzem:

- **Support (támogatás):** az asszociációs szabály helyesen képes előre jelezni a mintában található egyedei ennyi százalékára.
- **Confidence (bizalom):** az egyedek hány százalékára érvényes egy szabály
- **Significance (szignifikancia):** $\chi 2$ függetlenségi teszt eredménye, annak vizsgálatára, hogy statisztikailag kimutatható kapcsolat van egy szabály jobb

és bal oldala között. A null hipotézis esetén a bal és a jobb oldal függetlenek. Egy szabály elvethető – egy meghatározott szignifikancia szint mellett – amennyiben a khi-négyzet valószínűség kisebb, mint a szignifikancia szint.

- **Precision (pontosság):** százalékban kifejezett érték, egy szabály esetében azon egyedek aránya, amelyek helyesen lettek kiválasztva.

$$Pontosság = valós\ pozitív\ /\ (valós\ pozitív + ál\text{-}pozitív)$$

- **Recall (visszahívás):** azon helyes értékek százaléka, amelyeket a szabály kiválasztott

$$Visszahívás = valós\ pozitív\ /\ (valós\ pozitív + ál\text{-}negatív)$$

- **F-score (F-érték):** kombinált érték, amely a precision – recall közti viszonyt jellemzi (súlyozott harmónikus átlag) (Jurafsky & Manning, 2012)

$$F\text{-}score = (2 \times Precision \times Recall)\ /\ (Precision + Recall)$$

A kutatáshoz az adatokat *4.2 Az adatgyűjtő* alkalmazás című fejezetben leírtak szerint gyűjtöttem a Budapesti Corvinus Egyetem polgárairól. A kigyűjtött Facebook Like-okat az *5.4 Az* adatelemzés című fejezetben leírt pszichológiai API-val dolgoztattam fel és a kapott eredményeket elmentettem az adatbázisba a látogatók adataihoz. Az API csak megfelelő számú Facebook Like esetén képes személyes tulajdonságok jóslására. A 139 látogatóból összesen 95 látogatónak elegendő like-ja, az érvényes kimenethez. A jelen kutatáshoz az alábbi személyes jellemzőket mentettem el: Big5 (barátságosság, lelkiismeretesség, extrovertáltság, nyitottság, neurotikusság), IQ, élettel való megelégedettség, hit és politikai beállítottság.

6.1. Az adatok előkészítése

Az Apriori algoritmus kizárólag névleges (nominális) típusú változókat képes elfogadni inputként, ennek megfelelően a numerikus értékeket nominálissá kell konvertálni az algoritmus használata előtt. A névleges (nominális) mérési skála kizárólag az egységekhez rendelt számértékek egyező vagy különböző voltának jellemzésére alkalmas. Az adatbázisból kiexportált azon mezők esetében, ahol az értékek pusztán kódszámoknak tekinthetőek, ez az átalakítás pusztán formalitás. (Hunyadi & Vita, 2006)

Az alábbi lista tartalmazza, hogy milyen átalakításokat végeztem az adatokon:

- **Binárissá alakítás**
    - Ha az input változó értéke 0, akkor az output értéke 0, minden más esetben az output értéke 1
    - Átalakított nominális változók: is_DNT_on[28], has_flash, is_flash_blocked, has_silverlight, has_quicktime, has_java, has_PDFReader, has_AdobeReader, activity_dawn, activity_forenoon, activity_noon, activity_afternoon, activity_evening, activity_night
- **Binárissá alakítás vágással**
    - Ha az input változó értéke 0,5-nél kevesebb, akkor az output értéke 0, más esetben az output értéke 1
    - Átalakított nominális változók: s_yes, r_christian, b_agreeableness, b_conscientiousness, b_openness, b_neuroticism, b_extraversion, p_uninvolved, p_conservative, female
- **Kategóriaváltozóvá alakítás**
    - A Weka Discretize szűrőjét alkalmazva az input változó értékeit különböző kategóriákba soroltam, az inputváltozó mögötti zárójelben látható a létrehozott kategória száma
    - Átalakított ordinális skálán értelmezett változók: connection_type (2), font_counter (2), number_of_pageloads (3), cardinality_of_sessions (6), cardinality_of_locations (2), satisfaction_life (3), intelligence (3)
- **Nem változtatott**
    - Nem szükséges átalakítás, mivel az SQL lekérdezés végeredménye az Apriori algoritmus számára feldolgozható
    - Változók: browserFamily, osFamily, device, mobile_os_family, screen_size, geo_category, course, department

6.2. Elemzés

6.2.1. Klasszifikációs algoritmusok

Az osztályváltozó előrejelzésének pontosságát megjósolandó az alábbi klasszifikációs algoritmusokat alkalmaztam a mintára:

[28] A változók értelmezése a 8.7.5 fejezetben olvasható

- **ZeroR:** Nincs előrejelzési képessége, mert nem használ fel egyetlen attribútumot sem, pusztán viszonyítási alapként használatos. Egy feltétel nélküli szabály, amely ítéletrészében a leggyakoribb osztály áll. Az algoritmus összeállít egy gyakorisági táblázatot, majd kiválasztja a leggyakoribb értéket. (Bodon, 2010)
- **OneR:** A legegyszerűbb osztályozó algoritmus. Kiválaszt egy attribútumot és az osztályozásban kizárólag ezt használja. Annyi szabályt állít elő, ahány értéket felvesz a kiválasztott attribútum a tanítóhalmazban. (Bodon, 2010)
- **NaiveBayes:** A Bayes-tételen alapszik, mely az egyes attribútumok közötti függetlenséget feltételez. Egyszerűsége ellenére, széles körben elterjedt és meglepően jól teljesítő algoritmus. (Bodon, 2010)
- **J48:** Döntési fa algoritmus, az ID3 (Iterative Dichotomiser 3) algortimus implementációja. (Bodon, 2010)
- **RandomForest:** A véletlen erdők hatékony osztályozó algoritmusok, amelyek döntési fákat használva nagy adathalmazokkal is megbirkóznak. (Szabó, 2010)

Minden klasszifikációs algoritmus futtatása előtt kiválasztottam egy osztályváltozót, a többit pedig eltávolítottam a mintából, ezzel biztosítva, hogy azok a létrehozott szabályokban ne szerepeljenek.

	Kiválasztott osztályváltozó						
Klasszifikációs algoritmus	Élettel való megelégedettség	Intelligencia	Big5 - barátságosság	Big5 - lelkiismeretesség	Big5 - nyitottság	Big5 - neurotikusság	Big5 - extrovertáltság
ZeroR	85,3	85,3	67,4	73,7	69,5	65,3	67,4
OneR	82,4	82,4	65,3	70,5	68,4	60,0	64,2
Naïve-Bayes	79,4	76,5	55,8	69,5	60,0	50,5	54,7
J48	85,3	85,3	66,3	65,3	69,5	63,2	63,2
RandomForest	85,3	85,3	58,9	68,4	61,0	53,7	57,9

26. táblázat: az alkalmazott osztályozó algoritmusok által helyesen osztályozott változók százalékos aránya

A 26. táblázat eredményeiből kitűnik, hogy a ZeroR osztályozó algoritmus több esetben is felülmúlja a kifinomultabb algoritmusokat. Az élettel való elégedettség és az intelligencia jelezhető előre a legpontosabban a vizsgált osztályváltozók közül, emiatt a kutatás további részeiben csak ezekkel a változókkal fogok foglalkozni.

Az élettel való megelégedettség és az intelligencia esetében is a precision (pontosság) értéke 0,728, a recall (visszahívás) értéke 0,853 és az F-érték pedig 0,785 a legjobban teljesítő algoritmusok esetében (ZeroR, J48, RandomForest).

Osztályváltozó	Klasszifikációs algoritmus	Ál-pozitív arány	Hamis-pozitív arány	Pontosság (precision)	Visszahívás (recall)	F-érték
Élettel való megelégedettség	ZeroR	0,853	0,853	0,728	0,853	0,785
	OneR	0,824	0,855	0,724	0,824	0,770
	Naïve Bayes	0,794	0,859	0,720	0,794	0,785
	J48	0,853	0,853	0,728	0,853	0,785
	RandomForest	0,853	0,853	0,728	0,853	0,785
	RandomTree	0,794	0,859	0,720	0,794	0,755
Intelligencia	ZeroR	0,853	0,853	0,728	0,853	0,785
	OneR	0,824	0,858	0,724	0,824	0,770
	Naïve Bayes	0,765	0,868	0,715	0,765	0,739
	J48	0,853	0,853	0,728	0,853	0,785
	RandomForest	0,853	0,853	0,728	0,853	0,785
	RandomTree	0,735	0,873	0,711	0,735	0,723

27. táblázat: az intelligencia és az élettel való megelégedettség változókra alkalmazott klasszifikációs algoritmusok kimenete

A 27. táblázatból kitűnik, hogy a ZeroR, a J48 és a RandomForest algoritmus képes a legpontosabban előre jelezni az élettel való elégedettséget és az intelligenciát. Ezen algoritmusok a mintában szereplő példányok 85,3%-ára helyes értéket adtak eredményül.

6.2.2. Asszociáció

A begyűjtött 95 elemű mintára lefuttattam az Apriori algoritmust, annak érdekében, hogy szabályokat találjak a látogatókra jellemző személyes adatok és a személyre nem jellemző paraméterek között.

Intelligencia

Az algoritmus 18 ciklus után találta meg az elemzés további részéhez szükséges nagy adathalmazokat. A futtatások alkalmával a támogatottság (support) és a bizonyosság (confidence) értékei rögzítettek voltak.

		Támogatottság/support			
		0,1	**0,2**	**0,3**	**0,4**
Bizonyosság minimális szintje	**0,1**	0,34	0,34	0,31	-
	0,2	0,34	0,34	0,31	-
	0,3	0,34	0,34	**0,31**	-
	0,4	0,45	**0,45**	-	-
	0,5	0,54	-	-	-
	0,6	**0,67**	-	-	-
	0,7	-	-	-	-

28. táblázat: a talált szabályok bizonyossági (confidence) szintje rögzített bizonyosság és támogatottsági szint mellett

A 28. táblázatból kiolvasható, hogy a támogatottsági szint növekedésével a bizonyossági szint egyre csökken, tehát a megtalált szabályok a mintában szereplő egyének 10%-ának intelligenciáját képesek 67%-ban helyesen előre jelezni. Az algoritmus 768 szabályt talált 0,67-es bizonyossági szinten, amelyből mind szignifikáns. Ebből a két legrövidebb szabály az alábbiakban olvasható:

- **Szabály 1**
  - Windows 7 operációs rendszert használt
  - A képernyő szélessége nagyobb, mint 1024, de kevesebb, mint 1280
  - Egynél több földrajzi helyről érte el a weblapot
  - Legalább egyszer érte el a weblapot 18:00-23:00 óra között
  - Quicktime kiegészítő telepítve volt a látogató gépére
  - A használt böngészőnek volt PDF olvasási képessége
- **Szabály 2**
  - A képernyő szélessége nagyobb, mint 1024, de kevesebb, mint 1280
  - Egynél több földrajzi helyről érte el a weblapot
  - Legalább egyszer érte el a weblapot 18:00-23:00 óra között
  - Quicktime kiegészítő telepítve volt a látogató gépére
  - Silverlight kiegészítő telepítve volt a látogató gépére
  - A használt böngészőnek volt PDF olvasási képessége

Mindkét esetében az intelligencia értéke 0,51, ha a látogatóra igazak a felsorolt szabályok. A függelék 8.7.5 fejezetében olvasható képlet segítségével az IQ értéke 100,68.

Élettel való elégedettség

Az algoritmus ebben az esetben is 18 ciklus után találta meg nagy adathalmazokat, a támogatottság (support) és a bizonyosság (confidence) értékei szintén rögzítettek voltak.

		Támogatottság/support			
		0,1	**0,2**	**0,3**	**0,4**
Bizonyosság minimális szintje	**0,1**	0,34	0,34	0,31	-
	0,2	0,34	0,34	0,31	-
	0,3	0,34	0,34	**0,31**	-
	0,4	0,43	**0,44**	-	-
	0,5	0,54	-	-	-
	0,6	0,71	-	-	-
	0,7	**0,71**	-	-	-
	0,8	-	-	-	-

29. táblázat: a talált szabályok bizonyossági (confidence) szintje rögzített bizonyosság és támogatottsági szint mellett

Az algoritmus 1024 szabályt talált a 0,71-es bizonyossági szint mellett, amelyek mindegyike szignifikáns volt. A legrövidebb szabály esetében az osztályváltozó értéke 0,32 vagy annál nagyobb:

- A képernyő szélessége nagyobb, mint 1024, de kevesebb, mint 1280
- Legalább egyszer érte el a weblapot 7:00-10:00 és 18:00-23:00 óra között
- Quicktime kiegészítő telepítve volt a látogató gépére
- A használt böngészőnek volt PDF olvasási képessége

6.3. Összegzés

A már korábban ismertetett a Budapesti Corvinus Egyetem polgárairól gyűjtött mintában azt vizsgáltam, hogy lehetséges-e személyes tulajdonságokra vagy jellemzőkre következtetni a böngészés során elérhető nem személyes adatokból. A mintában kimutatható összefüggés található a böngésző személyisége és az általa használt szoftver és hardver környezet tulajdonságai, valamint a látogató online viselkedése között.

Gyakori elemhalmazok részhalmazait kerestettem az adatbányász körökben népszerű Apriori algoritmussal, amelyek közül a magas konfidenciaszinttel és támogatottsággal rendelkezőket vizsgáltam. (Agrawal & Srikant, 1994) Az algoritmus talált olyan szabályokat, amely a látogatók 10%-a esetében 67%-os konfidenciaszinten képes az intelligenciájukat, 71%-os konfidenciaszinten pedig az élettel való elégedettségüket előre jelezni.

A talált szabályok nagy valószínűséggel csak a vizsgált mintára érvényesek, de megfelelően képesek a webes adatok adatbányászati algoritmusokkal történő potenciál érzékeltetésére.

7. UTÓHANG

„… az internet, épp ahogy feltárta a világot mindenki számára, ugyanúgy tár fel mindnyájunkat a világ számára. És egyre inkább, a magánéletünk az ár, amit az összekapcsoltság miatt kell fizetnünk."

Gary Kovacs, TED (Gary, 2012)

A dolgozatom alapvetően két részre osztható: az első, leíró részben az anonimitás fogalmát és a látogatók azonosításának és követésének módszereit járom körül bemutatva a témával kapcsolatos szakirodalmat. A webes böngészés alapjaitól kezdve rendszerezve írom le a látogatók azonosításának és követésének lehetséges módszereit. A második részben a témával kapcsolatos kutatásaimat írom le:

- *Az eszközböngészőkről begyűjthető adatok jellemzői* című kutatásban az egy domain alól elérhető weboldalak számára elérhető paraméterek csoportjait vizsgáltam, amelyek közül az egy átlagos munkamenet ideje alatt állandó paraméterek járultak hozzá legnagyobb mértékben a látogató beazonosításához. Tapasztalataim szerint minél több változót használok a felhasználó azonosításához, annál nagyobb vizsgálatba bevont változók együttes bizonytalanság eloszlató képessége. Ez azt jelenti, hogy egy ismert sokaság elemei közül nagy bizonyossággal be tudjuk azonosítani az egyén által használt eszközböngészőt, feltéve, hogy a munkamenet ideje alatt állandó paraméterei ismertek. Amennyiben meg szeretnénk nehezíteni a követőink dolgát, célszerű proxy-n keresztül minél kevesebb paramétert megosztani a követő weboldalak és harmadik felek számára.
- *Az egyetemi polgárok pszichológiai jellemzői adatvédelmi szempontból* című kutatás során a – jelenleg legnépszerűbb közösségi oldal – a Facebook Like-jait elemeztettem a University of Cambridge – The Psychometrics Centre munkatársai által kifejlesztett pszichológiai API segítségével. Az API az input adatokból képes érdeklődési kört, politikai beállítottságot, hitvallást, családi állapotot, szexuális orientációt és különféle személyiségi jellemzőket kinyerni. A saját és a myPersonality Project keretében gyűjtött mintán különféle klaszterképző eljárásokat futtattam, amelyek stabil klasztereket eredményeztek. A különböző forrású klaszterek közül 2 nagyon hasonló

található: az egyedülálló, de nyitott, neurotikus és extrovertált tartalommegosztók és a komoly kapcsolatban lévő, kompromisszumképes, lelkiismeretes egyének.

- A *Személyes információ kinyerése webes adatokból* kutatásban szintén a Budapesti Corvinus Egyetem polgárairól gyűjtött mintában vizsgáltam a személyes tulajdonságok és a személyhez nem köthető, elérhető paraméterek közötti kapcsolatot. Kimutatható összefüggés található a böngésző egyén személyisége és az általa használt szoftver és hardver környezet tulajdonságai között. Ehhez a magas konfidenciaszinttel és támogatottsággal rendelkező gyakori elemhalmazok részhalmazait kerestettem az adatbányász körökben népszerű Apriori algoritmussal. (Agrawal & Srikant, 1994) Az algoritmus talált olyan szabályokat, amely a látogatók 10%-a esetében 67%-os konfidenciaszinten képes az intelligenciájukat és 71%-os konfidenciaszinten pedig az élettel való elégedettségüket előre jelezni.

Kutatásaim során egyaránt vizsgáltam a nem professzionális, **egy domain alól elérhető weboldalak** és professzionális a **közösségi oldalak** vagy **hirdetési ügynökségek** által hozzáférhető adatokat. A kutatásaimból és a feldolgozott irodalomból levont konzekvenciákat a böngészés kiemelt aktorai számára az alábbiakban összegzem:

- **Weboldalakat látogató egyének:** az egy domain alól elérhető weboldalak számára viszonylag kevés információ érhető el, ebből korlátozottan, de lehetséges személyes jellemzők kinyerésére, a kinyilvánított preferenciák miatt a közösségi oldalak és a hirdetési ügynökségek által hozzáfért adatokból lehetséges a személyes jellemzőkre következtetni. Minél pontosabban ismert a látogató személyes jellemzői, annál jobban a látogatóhoz illeszkedő tartalmakat és reklámokat lehetséges küldeni számára, ami a „filter bubbles” jelenségéhez vezet. Ezt elkerülendő célszerű tudatosan használni a közösségi oldalakat, keresőmotorokat és minden olyan weboldalt, amelybe a hirdetési ügynökségek reklámhordozókat ágyaztak be.
- **Adatvédelmi hatóság:** az Európai Unió kiberbiztonsággal foglalkozó felelős szervezete a „European Union Agency for Network and Information Security” (ENISA), amelyet 2004-ben hoztak létre annak érdekében, hogy ajánlásokat

tegyen, valamint a politikai szempontok kialakításánál és bevezetése során kulcsszerepet vállaljon az Európai Unió számára. (ENISA, 2017)

- **Szoftverfejlesztő/CIO**: A felhasználói/látogatói viselkedés és preferenciákból történő azonosítás megelőzése miatt érdemes az alkalmazások felkészítése a tárolt információk fokozott védelmére. Fontos az ENISA által kiadott „privacy by design" elv és a rá épülő keretrendszerek (ISO/IEC 29100) alkalmazása, amelynek célja az alkalmazások azok tervezési fázisában az információ védelmére való felkészítése. (Danezis, és mtsai., 2015)

A közösségi hálózatok megreformálták az információ terjesztését, hiszen használatukkal személyre szabott oldalon jelenik meg az ismerőseink által közzétett, az általunk kedvelt márkát és személyek által közzétett tartalmak. A Facebook népszerűségét jelzi, hogy a teljes emberi populáció közel 30%-a csatlakozott már hozzá és az online megjelenő cégek és szervezetek mára már nem a saját weboldaluk internetes címével hirdetik magukat, hanem a közösségi oldalon használatos elérhetőségeikkel.

„Minden klikkeléssel és minden érintéssel a képernyőn, olyanok vagyunk, mint Jancsi és Juliska és morzsákat hagyunk, személyes információ formájában, amint átkelünk a digitális erdőn keresztül."

Gary Kovacs, TED (Gary, 2012)

Az online életünket körülvevő közösségi hálózatok rengeteg adatot gyűjtenek a webet használókról. Egyelőre szimbiózis tapasztalható a webet használó látogatók és a szolgáltatásaikat ingyen áruba bocsátó webhely tulajdonosok közt, és csak néhány esetben lehetett visszaéléseket tapasztalni. Mivel a korlátozó jogszabályok egyelőre még gyerekcipőben járnak, valamint az online alkalmazások jelentős része nincs megfelelően biztosítva adatszivárgás ellen, javasolt körültekintően eljárni az adataink megosztása során.

8. FÜGGELÉK

8.1. HTML

A weboldalak megjelenítését leíró nyelv a **HTML** azaz **Hypertext Markup Language** első verzióját Tim Berners-Lee alkotta meg a ’90-es évek elején, de olyan gyorsan fejlődött, hogy 1997-ben már a HTML 4.0-ás verziójával találkozhatunk, majd 2008-ban publikálták a mérföldkőnek számító tervezetét, a HTML5-öt. 2012. decemberében W3C általi ajánlássá vált, majd 2014. októberében lett szabvány. Több új szintaktikai és szemantikai elemet tartalmaz és célja, hogy a desktop és a webes alkalmazások közti rést csökkentse.

A HTML kód legelső sora a DTD, (document type definition - dokumentum típus definíció) amely abban segít a böngésző számára, hogy az adott dokumentumot hogyan kell megjeleníteni. A DTD egy gép által olvasható nyelvtan, mely megadja a dokumentumban használható tag-ek listáját, valamint ezen tag-ekre érvényes szabályokat. Többek között a HTML alábbi verziói terjedtek el:

- HTML 4.01 (strict, transitional és frameset)
- XHTML 1.0 (strict, transitional és frameset)
- XHTML 1.1
- HTML5

A megfelelő DTD és az általa definiált szabályok használatával biztosíthatjuk, hogy a böngészők a tervezettnek megfelelően jelenítik meg a weboldalt.

8.2. HTTP lekérdezés

A szerveren található erőforrásokat URL-ek segítségével címezhetjük meg. Az URL a **Uniform Resource Locator** rövidítése és az erőforrások címzésére szolgáló protokoll, mely a 22. ábra, *URL szerkezete* látható módon épül fel. Az URL az URI (Uniform Resource Identifier – egységes erőforrás azonosító) családjába tartozik.

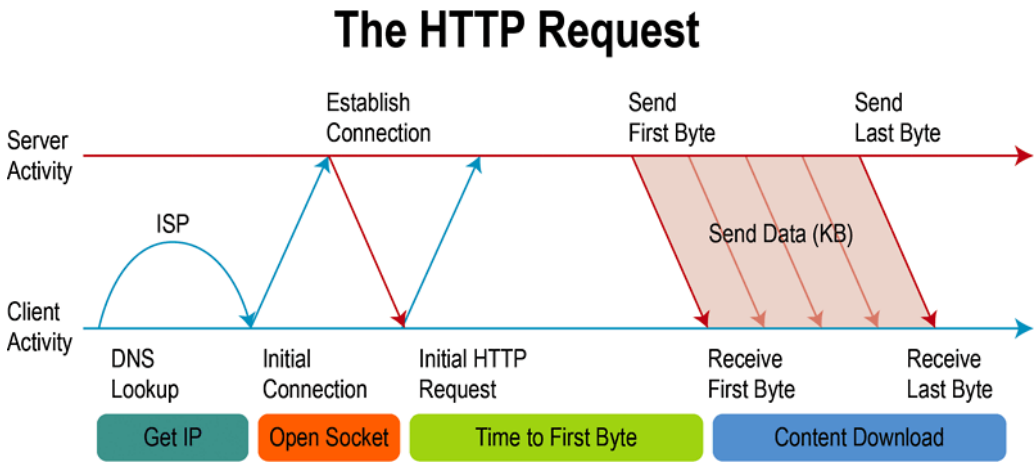
```
scheme://username:password@domain:port/path?query_string#fragment_id
```

22. ábra, URL szerkezete (saját szerkesztés)

Az **URL a** lekérdezni kívánt erőforrás elérésének meghatározására használt cím. Tartalmazza a lekérdezéshez használt protokollt, a szerver IP címét vagy domain-jét, a kommunikációhoz használt port számát, a lekérdezni kívánt erőforrás szerveren belüli helyét és a query stringet. A query string az URL-ben a ? után következő rész

adatok átküldésére szolgál, kulcs-érték párok formájában, egymástól & jellel elválasztva. Végül a # jel az oldalon belüli pozícionálás miatt szükséges.

A DNS feloldást követően a böngésző HTTP lekérdezéseken keresztül kommunikál a webszerverrel. A HTTP protokoll kapcsolat nélküli, ez azt jelenti, hogy a lekérdezés elküldését követően lecsatlakozik a szerverről és vár a válaszra.



23. ábra, A HTTP lekérdezés menete (Websiteoptimization.com, 2009)

A szerverről letölteni kívánt erőforrások lehetnek statikusak vagy dinamikusak. A statikus erőforrások tipikusan file-ok a szerveren, melyek tartalma nem vagy csak ritkán módosul. A dinamikusan előállított tartalmak előállításához a webszerver általában egy külső programozási nyelvhez fordul, ami minden alkalommal előállítja azt (kivétel persze, ha egy dinamikusan előállított tartalmat cache-elünk), majd visszaadja azt.

Amennyiben a HTTP lekérdezés egy erőforrásra hivatkozik, a szerver az alábbiak szerint járhat el:

- amennyiben a lekérdezni kívánt erőforrás egy a szerveren tárolt statikus file-ra vonatkozik, akkor a HTTP válaszban a szerver visszaküldi azt a kliens számára
- ha a kérés egy dinamikusan létrehozott HTML állományra hivatkozik, akkor a webszerver egy szerveroldali nyelvhez fordul, amelynek a outputja lesz az a HTML tartalom, amelyet a kliens számára visszaküldve annak böngészője megjelenít.

A letölteni kívánt állomány/erőforrás állapotáról a webszerver hibakódok segítségével kommunikál a klienssel, amely a HTTP protokoll része. Néhány gyakran előforduló HTTP hibakód:

- **200 „Success”:** sikeresen teljesítve, a válasz üzenet törzsében található a szerver által generált output
- **301 „Moved permanently”:** a kívánt erőforrást áthelyezték, a válaszból megtudható, hogy hova
- **404 „Not found”:** a kívánt erőforrás nem található
- **500 „Internal server error”:** általános hibaüzenet, többnyire a szerveren futó meghívott alkalmazás hibát dobott

A HTTP protokoll többféle metódussal teszi lehetővé egy állomány lekérdezését. A weblapok és online űrlapok túlnyomó többsége a lentebb kifejtett GET és POST metódust használja az összes művelet elvégzésére, azonban a HTTP 1.1 protokoll több lekérdezési metódust (Dan, 2004) támogat:

- **GET:** adat lekérdezésére szolgáló metódus. A lekérdezés paraméterei a query stringben lesznek elküldve. Az URL legnagyobb megengedett hossza 1024 byte, így a query string emiatt nem lehet túl hosszú, valamint érzékeny adatok átküldésére nem célszerű használni.
- **HEAD:** ugyanaz, mint a GET, csak ebben az esetben a szerver csak a fejlécet küldi vissza (response header)
- **POST:** adat küldésére szolgáló metódus. A lekérdezés testében (body) van a szerverre küldött adat. (szöveg, adatállomány stb.)
- **PUT:** adat megváltoztatására vagy létrehozására szolgáló metódus
- **DELETE:** adat törlésére szolgáló metódus
- **OPTIONS:** kommunikációval kapcsolatos információk lekérdezésére szolgál
- **TRACE:** visszaküldi a kliens számára az elküldött üzenetet. Annak tesztelésére találták ki, hogy a kliens és a szerver között lévő csomópontok mit változtatnak meg az elküldött adatcsomagban.
- **CONNECT:** TCP/IP tunnel-lé változtatja a kommunikációs kapcsolatot, így lehetőség nyílik HTTPS kapcsolatra HTTP-t támogató proxy-n keresztül is.
- **PATCH:** az elküldött adatok csak egy része a rekordnak, módosításra használható

A GET és a POST metódus idempotens, a HEAD, OPTIONS, GET és a TRACE metódus biztonságos, mivel az nem végez módosítást a szerveren.

GET	POST
GET /blog/?name1=value1&name2=value2 HTTP/1.1 Host: carsonified.com	POST /blog/ HTTP/1.1 Host: carsonified.com name1=value1&name2=value2

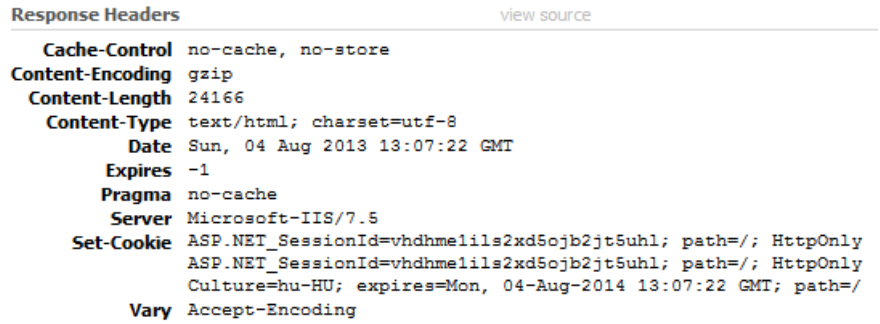
30. táblázat: A GET és POST lekérdezés HTTP fejléce (saját)

Az adatok tárolásánál 4 alapműveletet szoktak megkülönböztetni: létrehozás, olvasás, módosítás és törlés. Az adatbázis műveletek is eszerint csoportosíthatóak és a fent említett metódusok szintén megfeleltethetőek a 4 alapműveletnek az alábbiak szerint:

Művelet	SQL	HTTP
Create	INSERT	PUT/PATCH
Read (retrieve)	SELECT	GET
Update (Modify)	UPDATE	POST
Delete (Destroy)	DELETE	DELETE

31. táblázat: az SQL és a HTTP CRUD műveletei (Medic, 2014)

Miután a böngésző megkapta a választ a szervertől, elkezdi annak feldolgozását. Ha a HTTP válasz fejlécében érkezik süti, akkor azt a böngésző létrehozza. A 24. ábra a WizzAir.hu webhely lekérdezésére irányuló HTTP válasz látható. A Set-Cookie rész első két sorában látható két munkamenet azonosító, majd látható a Culture nevű változóba elmentették a lokalizációt. A süti lejárati ideje 1 év és az egész webhelyre érvényes.



24. ábra, a WizzAir.hu főoldalának HTTP válasza (saját)

Ha a feldolgozandó adat egy HTML állomány, akkor értelmezi azt, letölti a külső hivatkozásokat, majd elkezdi kirajzolni a dokumentumot és ahol Javascriptet talál, azonnal elkezdi futtatni, ez akkor is megtörténik, ha a dokumentum bizonyos részei még nem töltődtek le. Erre a problémára a megoldás, ha figyelünk valamilyen betöltődés/feldolgozás vége eseményre.

A dokumentum rajzolásának végén két fontos esemény hívódik meg:

- **onDomReady:** akkor hívódik meg, ha a memóriában már rendelkezésre áll a DOM, tehát bármelyik része hivatkozható, módosítható és törölhető, viszont a külső erőforrások (CSS, képek, iframe-ek tartalma stb.) még nem feltétlenül állnak rendelkezésre
- **onLoad:** akkor hívódik meg, ha az egész dokumentum letöltődött a DOM rendelkezésre áll a memóriában, valamint letöltődött az összes külső erőforrás is

Az alap Javascriptben csak az onLoad esemény létezik az onDomReady nem, így azt csak különböző library-k vagy framework-ök segítségével lehet csak elérni. A Javascript futtatásához nem szükséges megvárni, hogy az összes kép vagy stíluslap letöltődjön, ha a memóriában rendelkezésre áll a DOM, el lehet kezdeni a Javascript futtatását.

Az **AJAX** az **Asynchronous Javascript and XML** rövidítése, amely webes technológiák csoportját foglalja magába. Tulajdonképpen adatok küldése és fogadása a szervertől az oldal újratöltése nélkül. A kérést Javascript indítja el a háttérben és aszinkron[29] módon vár a válaszra.

8.3. Javascript

Jelenleg a **Javascript**/ECMAScript[30] az egyetlen, amelyet az összes böngésző támogat. A nyelvet eredetileg a webböngészőkbe tervezték, hogy elősegítse a RIA-k létrejöttét, (rich internet application) interaktívvá tegye a weblapokat vagy képessé tegye a böngészőt aszinkron kommunikációra.

A **DOM (Document Object Model)** egy platform és nyelvfüggetlen konvenció, HTML, XHTML és XML dokumentumokkal való interakcióra és adatok

[29] a kérés elküldése és a válasz visszaérkezése között eltelt idő alatt is fut tovább az alkalmazás, azaz nem akadályozza annak futását

[30] A Netscape kezdeményezésére 1997-ben fogadta el az ECMA (European Computer Manufacturers Association) ECMAScript néven

megjelenítésére alkalmas. A DOM fa struktúrában tárolt és az egyes elemek a saját függvényeikkel címezhetőek és manipulálhatóak. Ez tulajdonképpen a böngésző API-ja.

```
<script type="text/javascript">
      window.location.href = 'http://google.com';
</script>
```

25. ábra, HTML DOM objektum kezelése Javascript-tel (saját szerkesztés)

8.4. Süti (cookie)

A **süti (cookie)** egy kliensoldali adattárolási módszer neve. A böngésző a kliens oldalon egy kódolatlan szöveges állományban kulcs-érték párok formájában tárolja el a kívánt adatokat. Az eltárolható sütik maximális mérete 4 kB, de ez az érték böngészőnként eltérhet, valamint domain-enként maximálisan 50 sütit lehet eltárolni. A kliens hardverére telepített különböző böngészők nem képesek olvasni és írni egymás sütijeit. Ez azért fontos tényező, mert egy felhasználó által használt hardveren a különböző böngészők nem férnek hozzá egymás sütijeihez, így ha a sütiket a felhasználó követésére használjuk, ha a felhasználó böngészőt vált, beazonosítása nehézkessé válik.

A sütik létrehozásánál megadható a lejárati idejük, amely lejárta után a süti a böngésző által már nem hozzáférhető. A süti törléséhez így csak egy múltbeli időpontra kell állítani a lejárati idejét. Amennyiben nem adunk meg lejárati időt, a süti a munkamenet lejártakor törlődik.

A süti létrehozásakor megadhatjuk annak láthatósági/hozzáférhetőségi körét. Alapesetben a süti mindig csak az adott domain és az URL-ben lévő elérési út számára elérhető. Szükség esetén feloldható ez a korlátozás oly módon, hogy az adott domain-en lévő más weboldal is hozzáférhessen. Természetesen egy adott domain alól más domain számára nem hozható létre süti.

A sütiket az alábbi módokon lehet létrehozni:

- a szerveroldali nyelv a **HTTP válaszban** küldi azt, a böngésző pedig létrehozza a kulcs-érték párokat tartalmazó file-t a kliens eszközén. A 26. ábra látható, hogy a HTTP válasz fejlécében a Set-Cookie paraméter tartalmazza a sütiben eltárolandó kulcs-érték párokat

26. ábra, szerver által küldött sütik (saját felvétel)

- a **kliensoldali programozási** nyelv utasítja közvetlenül a böngészőt a süti létrehozására. A 27. ábra látható, hogy az alap Javascripttel magunknak kell összeállítani a kulcs-érték pár, valamint a hozzá tartozó lejárati időt, míg jQuery esetében paraméterként adjuk át a kulcsot és az értéket, az extra paramétereket pedig objektumként

```
// Javascript
var exdate=new Date();
exdate.setDate(exdate.getDate() + exdays);
var c_value=escape(value) + ((exdays==null) ? "" : ";
expires="+exdate.toUTCString());
document.cookie=c_name + "=" + c_value;
// jQuery
$.cookie("test", 1, { expires : 10 });
```

27. ábra, süti létrehozása Javascripttel és jQuery segítségével (saját)

Az adott domain által létrehozott süti tartalmát a böngésző minden egyes lekérdezés során a HTTP fejlécben elküldi a szerver számára, így ha a sütiben sok adatot tárolunk, akkor azok minden egyes lekéréssel együtt el lesznek küldve a szerverre pazarolva a sávszélességet. Az esetek többségében (képek, CSS, Javascript állományok, statikus tartalmak stb.) ezek az adatok teljesen feleslegesen lesznek elküldve a szerverre. Ilyen esetekben szokták azokat az erőforrásokat, amelyekhez nincs szükség sütire, kitenni egy cookie mentes domainre, aldomainre vagy CDN-re[31].

A céljuk és a felhasználási módjuk alapján az alábbi sütiket különböztetjük meg:

[31] A Content Delivery Network (CDN) a webes tartalomelosztás egyik módszere. A lényege, hogy a tartalmat a Föld különböző pontjain helyezik el redundásan és az igény felmerüléséhez legközelebbi csomópont szolgál ki.

- **Munkamenet süti:** nincs beállítva a lejárati idejük, így csak addig léteznek, amíg a felhasználó be nem zárja a böngészőjét
- **Perzisztens/nyomkövető süti:** a munkamenet lejártát vagy a böngésző bezárását követően nem törlődnek a felhasználó gépéről, a lejárati idejük egy konkrét időpontra van beállítva. A beállított időpontig minden egyes lekéréssel a HTTP fejlécben el lesznek küldve a szerverre.
- **3rd party cookie/3. fél által készített süti:** általában hirdetők által alkalmazott technológia, az adott weboldal meglátogatása során a lekérdezett weboldaltól különböző domain alá elhelyezett süti. A sütik forrás domain-je nem egyezik meg a megtekintett oldal domain-jével, emiatt könnyű azokat kiszűrni. DNS aliasing technológiával a harmadik fél által használt domain egy DNS CNAME alias-szal átirányítható a meglátogatott oldal domain-jére, így azok kiszűrése nem lehetséges. (Chris, Ashkan, Nathaniel, & Dietrich, 2012)

A sütik kis méretük (maximálisan 4kB) ellenére igen sokoldalúan felhasználhatóak, többek között munkamenet kezelésére, a weboldal személyre szabásra és felhasználói élmény növelésére, valamint nyomkövetésre.

A törvény, amely a sütiket és más kliensoldali adattárolási technológiákat szabályoz 2011. május 26-án lépett hatályba Európában, (All About Cookies, 2011) kötelezi a weboldalak készítőit, hogy mielőtt sütiket mentenének a kliens böngészőjébe, előzőleg engedélyt kell kérni a felhasználótól. (opt-in)

A sütikben a weboldal tetszőleges adatot tárolhat. Érzékeny adatokat nem célszerű benne tárolni, mivel az kódolatlan állományként tárolódik a kliens gépén. A süti megtartja az értékét mindaddig, amíg az alábbi események valamelyike be nem következik:

- **A süti érvényessége lejár:** a lejárati dátuma korábbi, mint az internetezéshez használt eszközről kiolvasható időpont, ebben az esetben a böngésző a következő futása alkalmával az internetezéshez használt eszközről is törli a sütit tartalmazó állományt, így az a továbbiakban nem lesz olvasható.
- **A süti tartalma nem olvasható:** a böngésző számára nem hozzáférhető a süti tartalma, mert az file szinten nem hozzáférhető. Ennek az is lehet az oka, hogy a sütik tartalmát szándékosan törölték vagy olvashatatlanná tették az internetezéshez használt eszközről.

8.5. CSS3 és HTML5 képességek

- CSS3 képességek – Javascript segítségével detektálhatóak
    - @font-face, background-size, border-image, border-radius, box-shadow, flexible box model – különböző CSS3 képességek
    - hsla – hue-saturation-light-alpha színpaletta használata
    - több háttérkép használatának lehetősége
    - rgba – piros-zöld-kék-alpha színskála használatának lehetősége. A CSS korábbi verziói csak az RGB színskálát támogatták, az RGBA esetében lehetőség nyílik az áttetszőség megadására is. Ez akkor lehet hasznos, ha az aktuális elem mögött még van valamilyen többszínű vizuális elem.
    - text-shadow – szöveg árnyékolásának lehetősége
    - CSS animáció lehetősége – a kijelölt elemek animálásának lehetősége
    - hasábok létrehozásának lehetősége – lehetséges-e CSS-ből hasábokra bontani egy szöveget
    - generated content (:before/:after) – CSS segítségével a kijelölt elem elé illetve mögé lehet beszúrni tartalmat
    - CSS gradients – CSS színátmenet képzésének lehetősége
    - CSS tükröződés – egy elem tükröződésének beállításának lehetősége
    - CSS 2D transzformáció támogatása
    - CSS 3D transzformáció támogatása
    - CSS transitions/átmenetek (transitions) – az animáció és az átmenet között az a különbség, hogy amíg az animáció folyamatos, az átmenetet valamilyen eseményhez kötött, pl: kattintás stb.
- HTML5 képességek – szintén Javascript segítségével detektálhatóak
    - applicationCache – alkalmazás cache támogatása
    - canvas – grafikák rajzolásának lehetősége tag-ekkel vagy Javascriptből
    - canvastext – szöveg írása a festővászonra
    - Drag and Drop támogatása – drag and drop események támogatása, lehetőség van az asztalról file-ok beemelésére, valamint a böngészőből file-ok és tartalmak kiemelésére
    - hashchange esemény támogatása – akkor következik be, ha a meglátogatott URL hashtag mögötti része megváltozik

- history – history management támogatása
- audio – ogg, mp3, wav és m4a audiofile-ok lejátszásának támogatása
- video – ogg, webm és h264 formátumú videofile-ok lejátszásának támogatása
- indexedDB – indexedDB kliensoldali adattárolás támogatása
- beviteli mező attribútumok (autocomplete, autofocus, list, placeholder, max, min, multiple, pattern, required, step) támogatása
- beviteli mező típusok (search, tel, url, email, datetime, date, month, week, time, datetime-local, number, range, color) támogatása
- localStorage és sessionStorage támogatása
- postmessage – ablakok közti kommunikáció támogatása
- webSockets – socket kommunikáció támogatása
- webSQLDatabase – kliensoldali SQL adatbázis támogatása
- webWorkers – többszálúság támogatása
- geoLocation – geolocation támogatása
- SVG és inline SVG – Scalable Vector Graphics támogatása
- SVG clip paths – SVG rajzolás során alkalmazható clipPath, ami azt jelenti, hogy a rajzolási terület
- SMIL – Synchronized Multimedia Integration Language támogatása, multimédiás előadások készíthető vele
- touch – érintőképernyő támogatása
- webGL – rajzolásoz használt API támogatása

- A kliens oldal terhelését követően pedig a szerver oldal terhelését elemzem ugyanolyan szempontok szerint.

8.6. Látogatók azonosításához használható kiemelt paraméterek elemzése

Megnevezés	Előny	Hátrány
IP cím	• dinamikus IP cím esetében a munkamenet idejére azonosítja az eszközböngészőt • fix IP cím esetében akár	• a privát IP cím biztonsági okokból nem kérdezhető le a kliensoldalon • a szerveroldalon hozzáférhető IP cím nem feltétlenül a kliens valós IP címe, hiszen a kliens lehet egy proxy vagy egy

	hosszabb időn keresztül, több munkamenet idejére azonosítja az eszközböngészőt	router mögött is, amely elfedi az IP címét • geoLocation hiányában az IP címből lehet feloldani földrajzi pozíciót
Képernyő magassága / szélessége és bitmélység	• eszközökre jellemző a felbontás visszaellenőrzésh ez használható • a munkamenet közben történő képernyőváltással megismerhető, hogy a felhasználó milyen képernyőkhöz fér hozzá	• a felhasználó munkamenet közben is válthat képernyőt, így ezek az adatok menet közben is változhatnak
Aktuális böngészőablak mérete		• felhasználó akármikor megváltoztathatja, felesleges lementeni
Useragent string	• minimum a munkamenet idejére azonosítja az eszközböngészőt • értékes információval szolgál a kliens böngészőjéről, operációs rendszeréről és hardveréről	• a Useragent string sűrűn változhat, akár havonta is frissítheti a böngészőjét a felhasználó
HTML5 & CSS3 képességek	• UX szempontból lehet érdekes	• böngésző típus és verzióra jellemző
Beépülő modulok jelenléte és verziószáma	• a beépülő modulok listája és verziója egyedi lehet	
Telepített betűtípusok	• a telepített betűkészletek listája egyedi lehet	• Flash beépülő modul nélkül nem lehet lekérdezni

	• inputként szolgálhat a telepített szoftverek listájához	• Flash blokkoló alkalmazás jelenlétével nem állítható elő
Hálózat típusa	• értékes információval szolgál a felhasználó internetkapcsolatáról	• csak mobil eszközök esetében elérhető, számítógép esetén UNKNOWN értékkel tér vissza • nem detektálható ha az adott eszköz tethering[32]-gel kapcsolódik az internetre
Felhasználó böngészési előzménye	• a böngészőből biztonsági okok miatt nem érhető el • szerveroldalon a felhasználó elmentett felhasználó böngészési előzményeiből lekérdezhető	• megtudható belőle, hogy a felhasználók a webhely mely részeit nézik gyakrabban
Oldalletöltés ideje	• Kliens és szerveroldalon is detektálható • Minden elmentett rekordhoz kell társítani	
Generált eszközböngésző azonosító	• egyértelműen azonosítja az eszközböngészőt	• a felhasználó módosíthatja vagy törölheti
Billentyűzet és egérhasználat	• Felhasználóra jellemző	• körülményes az azonosításhoz szükséges adat begyűjtése • különböző eszközböngészőkön jelentősen eltérhet • az adatok elemzésből személyiségi jogokat sértő eredmények is születhetnek

32. táblázat: kiemelt paraméterek elemzése

[32] internet megosztása mobil eszközön keresztül

8.6.1. HARDVER JELLEMZŐI

- **Hálózat típusa** (ethernet, 3G, Wifi stb.): mobil kommunikációs eszközök esetében Javascript segítségével elérhető
- **Eszköz típusa, verziója:** szintén a HTTP lekérdezés fejlécében található User-agent string-ből kinyerhető
- **Képernyő szélessége és magassága**, megjelenítéshez használt **bitmélység:** csak a kliensoldalon áll rendelkezésre, Javascript segítségével lekérdezhető

8.6.2. OPERÁCIÓS RENDSZER, SZOFTVERKÖRNYEZET JELLEMZŐI ÉS HÁLÓZATI JELLEMZŐK

- **Csatlakozáshoz használt ISP adatai**: online adatbázisok adatbázisokból kinyerhető információ, sok esetben kikövetkeztethető belőle a kapcsolat típusa is, ha tudjuk, hogy az adott ISP milyen kapcsolatot szolgáltat. Van olyan szolgáltató, amely csak vezetékes internetet és van olyan is, amelyik csak mobilinternetet szolgáltat.
- **Operációs rendszer típusa és verziószáma**: a HTTP lekérés fejlécében lévő User-agent string-ből kinyerhető
- **IP cím:** a szerveroldalon környezeti változóként rendelkezésre áll
- **Host**: az IP címből reverse DNS-sel visszafordítható
- **geoLocation szélességi és hosszúsági fok, pontosság** – Javascript segítségével kérdezhető le a böngészőből, felhasználói beleegyezés szükséges
- **Telepített betűtípusok:** a kliensoldalon érhető el Flash kiegészítő használatával
- **Telepített alkalmazások:** a telepített betűtípusok jellemzőek lehetnek egy-egy alkalmazásra, a telepített betűtípusok elemzéséből következtetni lehet a látogató gépére telepített alkalmazásokra. Problémát jelenthet az alkalmazások telepítésekor feltelepített betűtípusok listájának összegyűjtése, valamint a lista frissen tartása.

8.6.3. ESZKÖZBÖNGÉSZŐ JELLEMZŐI

- **Böngésző típusa és verziószáma**: szintén a HTTP lekérdezés fejlécében lévő User-agent string-ből kinyerhető

    - httpUserAgent – a böngésző és operációs rendszer adatait tartalmazó szöveg
    - httpAccept – a böngésző által támogatott állományok MIME típusa
    - httpAcceptEncoding – a böngésző által támogatott tömörítési típusok és prioritásuk
    - httpAcceptLanguage - a böngésző által támogatott nyelvek és prioritásuk
    - httpAcceptCharset - a böngésző által támogatott karakterkódolások és prioritásuk
    - Do-not-Track: a HTTP lekérdezés fejlécéből kinyerhető paraméter
- **CSS3 és HTML5 képességek** (lásd a Függelékben)
- **Beépülő modulok jelenléte és jellemzői**
    - Flash verziója, Flash blokkoló működik-e, Silverlight verziója, Quicktime verziója, DevalVR verziója, Shockwave verziója, Windows Media Player verziója, Java verziója, PDF olvasó van-e a böngészőben, Adobe PDF Reader verziója, VLC Player verziója, Firefox PDF reader verziója, RealPlayer verziója
- Felhasználó **böngészési előzménye** – a felhasználó böngészési előzménye, a böngészési előzmény nem hozzáférhető, azonban a szerveren környezeti változóként elérhető a hívó URL[33] (referrer) (Jia-Ching, Chu-Yu, & Vincent, 2012)
- **Inkognitó mód** használata: kliensoldalon keresztül volt érhető el Javascript segítségével, az új böngészőkben nem lehet detektálni
- **generált eszközböngésző azonosító** – a szerveroldalon a böngésző jellemzőiből generált adat, amely az eszközböngésző nyomkövetésére alkalmas, a kliensoldalon sütiben és amennyiben lehetséges, akkor localStorage-ben tárolódik, a szerveroldalon pedig célszerű adatbázisban eltárolni
- **Közösségi hálózatok által hozzáférhető adatok** – harmadik féltől hozzáférhető adat, amely a szerver és a kliens oldalon is lekérdezhető. A lekérdezést célszerű a szerveroldalra helyezni, mivel a túl sok adat egyidejű

[33] Amennyiben a nyomkövető kódot az oldalon belül egy iframe-ben helyezzük el, a referrer értéke a maga a meglátogatott weblap lesz

lekérdezése csökkentheti a böngészési élményt. Egyszeri elhasználói beleegyezés szükséges, és pl: a Facebook esetében időközönként meg kell újítani az engedélyt.

- OpenID
- Facebook azonosító és az összes Facebook-on tárolt adat
- Google+ azonosító és az összes Google+-on tárolt adat
- LinkedIn azonosító és az összes LinkedIn-en tárolt adat
- egyéb közösségi hálózatokon tárolt és lekérdezhető adat

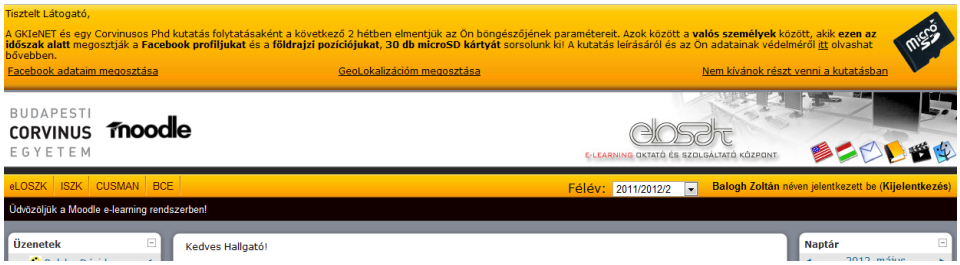
8.6.4. LÁTOGATÓ JELLEMZŐI

- egy meglátogatott **webhely bejárásának módja** – a lementett adatokból kinyerhető
- **egérmozgás** (x és y koordináta és időpont) – a kliensoldalon detektálható Javascript segítségével
- **billentyűzet leütés** (billentyű lenyomása és felengedése, időpont) – Javascript segítségével detektálható a kliensoldalon

A lista nem teljes, egy jelenkori állapotot tükröz, és jól szemlélteti, hogy a böngészőkből milyen mennyiségű adat hozzáférhető és ennek nagy része akár már 1 oldal meglátogatásával is hozzáférhető. A 2012 tavaszán elvégzett adatgyűjtési kísérletem során a felhasználók átlagosan 300 kB adatot osztottak meg magukról és egy oldal letöltésével több, mint 4 kB-ot.

8.7. Adatgyűjtő alkalmazás

A fejezet az általam fejlesztett adatgyűjtő alkalmazás dokumentációját tartalmazza. A *4.2 Az* adatgyűjtés fejezetben kifejtettem az alkalmazással szemben felállított követelményeimet, valamint a saját alkalmazás fejlesztésének indoklását. Miután az Budapesti Corvinus Egyetem Informatika Intézetének vezetésétől megkaptam az engedélyt, hogy az Egyetem e-learning rendszerébe beépítsem az adatgyűjtő alkalmazást felmértem az adatgyűjtő alkalmazás beépítésének lehetőségeit. Ennek eredményeképpen a bejelentkezést követően az alkalmazás egy iframe-en belül jelenik meg az oldal tetején.



28. ábra: Adatgyűjtő alkalmazás megjelenése a Corvinus e-learning rendszerében (saját felvétel)

Az alkalmazás szerveroldali részét egy az Egyetemtől igényelt tárhelyre töltöttem fel, amely az alábbi jellemzőkkel rendelkezett:

- 100 MB web tárhely, PHP 5 futtatási lehetőséggel
- 5 GB MySQL tárhely

Mivel az Egyetemtől kapott tárhely PHP szerveroldali alkalmazások futtatására alkalmas, ezért az adatgyűjtő alkalmazás szerveroldali részét ezen a programnyelven kellett megírnom. A kliensoldali részét HTML és Javascript nyelven írtam meg. Célom az volt, hogy az alkalmazás a böngészők legszélesebb körén futtatható legyen[34], amihez nagy segítséget nyújtott a Modeznizr Javascript library. A Modernizr detektálja a böngésző képességeit és egységes formában teszi elérhetővé a programozó számára.

8.7.1. Az alkalmazás működése

Az e-learning rendszerbe történő belépést követően minden egyes oldal tetején megjelent az adatgyűjtő alkalmazás, melynek a működése a *28. ábra: Adatgyűjtő alkalmazás megjelenése a Corvinus e-learning rendszerében* című tekinthető meg és az alábbiak szerint működik:

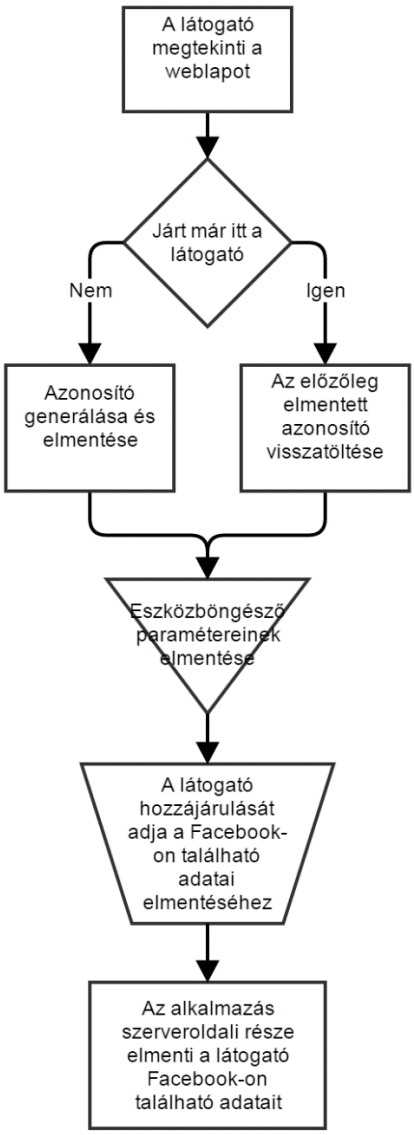
- Az oldal betöltődését követően az alkalmazás detektálta a böngésző képességeit, a beépített kiegészítők jelenlétét és verziószámát és a telepített betűkészletek listáját, majd elküldte a szerveroldal számára, ahol ez egy rekordként az adatbázisban került eltárolásra.
    - Amennyiben a látogató még nem járt korábban a meglátogatott weboldalon, akkor az alkalmazás szerveroldali része legenerált

[34] És a legnépszerűbb 3 böngészőben hibátlanul fusson (Google Chrome, Internet Explorer 6+, Mozilla Firefox 10+)

számára egy egyedi azonosító kódot, amelyet visszaküldött a kliens számára. Ez az egyedi azonosító a kliens oldalon sütibe és localStorage-ban került eltárolásra későbbi azonosítás céljából.

- Ha azonban a látogató korábban már járt a weboldalon, akkor a sütiből vagy a localStorage-ből visszatöltött az azonosításra alkalmas kódot a detektált adatokkal együtt visszaküldte a szerver számára, így azokat eszközböngészőhöz lehet kapcsolni.

- A **földrajzi pozíció** megosztása gombra történő kattintás esetén a kliens böngészője a 9. ábra*, a Chrome engedélyt kér a látogató földrajzi pozíciójához* látható módon rákérdez, hogy valóban szeretné-e megosztani a földrajzi koordinátáit a weboldallal. Amennyiben a felhasználó az igen gombra kattint, a kliens azt a háttérben elküldi a szerver számára a földrajzi koordinátáit egy lekérdezésben.
- A **Facebook-on található adataim megosztása** gombra történő kattintást követően egy felugró ablakban a Facebook oldalán engedélyezni kell a Facebook számára, hogy az alkalmazás hozzáférhessen a látogató adataihoz. Amennyiben a látogató megerősíti ezt, az alkalmazás egy párhuzamos szálon lementi a látogató Facebook-on található személyes adatait a látogató egyedi azonosítójával együtt
    - Abban az esetben, ha a látogató már egyszer engedélyezte az alkalmazás számára, hogy az aktuálisan használt eszközböngészőből a Facebook-on tárolt személyes adatait lementse, akkor többet nem jelenik meg az oldal tetején a *28. ábra: Adatgyűjtő alkalmazás megjelenése a Corvinus e-learning rendszerében* látható kérdés



29. ábra: Adatgyűjtő alkalmazás vázlatos működése (saját szerkesztés)

Amíg a kliensnél a **sütiben** vagy a **localStorage**-ban eltárolt egyedi azonosító hozzáférhető, a későbbi oldalletöltés adatait is hozzá lehet kötni a korábbi munkamenetekhez. Szándékosan tároltam el két különböző helyen a látogató azonosítóját, mert, ha az egyik azonosítót a felhasználó szándékosan kitörli, az a másikból visszaállítható.

8.7.2. Hozzáférhető adatok relevanciája és terhelése

Az adatgyűjtő alkalmazás tervezésénél kiemelt szempont volt, hogy a felhasználói élményt az alkalmazás nem rontsa. (pl: ne kelljen a látogatónak várnia, amíg az alkalmazás a háttérben a Facebook-on tárolt adatait elmenti) Az alábbi táblázatokban összefoglaltam a kliens és a szerver oldalon hozzáférhető adatok relevanciáját és kinyeréséhez szükséges hardver erőforrás igényét.

Kliens oldal

A táblázatban a kliensoldalon hozzáférhető paramétereket gyűjtöttem össze, amelyek Javascript segítségével érhetőek el.

Tényező neve	Hozzáférés módja	Hardver terhelése	Relevancia
IP cím	Nem hozzáférhető a kliensoldalon	-	-
képernyő szélessége, magassága és a megjelenítés bitmélysége	Javascript segítségével lekérdezhető	alacsony	magas
eszköz típusa és verziószáma	A szerveroldalon a lekérdezés fejlécében a User-Agent String-ből is kinyerhető	közepes	magas
csatlakozáshoz használt hálózat típusa	általában mobil eszközök esetében a kliensoldalon rendelkezésre álló érték	alacsony	közepes
HTTP fejléc	szerveroldalon környezeti változóként rendelkezésre áll	alacsony	magas
CSS3 képességek	Modernizr library segítségével detektálható a kliensoldalon	közepes	közepes
HTML5 képességek		közepes	közepes
beépülő modulok jelenléte és verziószámának detektálása	PluginDetect[35] library segítségével a kliensoldalon hozzáférhető	közepes	magas
geoLocation	felhasználói hozzájárulással a kliensoldalon hozzáférhető	alacsony	magas
felhasználó böngészési előzménye	előzőleg meglátogatott és a nyomkövető script által adatbázisba lementett	közepes	közepes

[35] http://www.pinlady.net/PluginDetect/All/

	előzmények a szerveroldalon lekérdezhetőek		
egérmozgás követése	a kliensoldalon detektálható	magas	közepes
billentyű leütések követése		magas	közepes
oldalletöltés ideje	kliens és szerveroldalon egyaránt detektálható	alacsony	magas
meglátogatott URL	a kliens és szerveroldalon egyaránt környezeti változóként elérhető	alacsony	magas
generált eszközböngésző azonosító	a szerver által generált azonosító, amelyet vissza kell juttatni a kliens számára, majd ott elmenteni sütibe és localStorage-ba	közepes	magas
közösségi hálózatok által hozzáférhető adatok	felhasználó hozzájárulása szükséges, harmadik féltől hozzáférhető adat	magas	magas
telepített betűtípusok listája	Flash plugin szükséges a lista letöltéséhez, majd Javascript segítségével kinyerni a Flash plugin-ből	közepes	magas
inkognitó mód	régebben volt lehetséges az inkognitó mód detektálására, a korszerű böngészőkben ez már nem lehetséges	közepes	alacsony

A táblázatban a detektálható paraméterek és a magyarázatuk mellett feltüntettem azok hardver terhelését és relevanciáját. A hardver terhelés oszlopban feltüntetett értékek az adott paraméter kinyeréséhez szükséges kliens oldali erőforrás szintjét jelzik. Az alacsony hardverigény egy környezeti változó kiolvasását jelenti, a közepes hardverigény a kliensoldalba betöltött middleware-től kapott értéket jelenthet,

a magas hardverigény pedig harmadik féltől származó adatok áttöltését vagy a billentyűleütések/egérmozgás folyamatos detektálását és elküldését jelenti.

Az alkalmazás tervezésénél az alacsony és közepes hardverigényű paraméterek mindegyikét elmentettem, a magas hardverigényűek közül a közösségi hálózatokról elérhető személyes adatokat a szerveroldalon töltöm át, így az nem befolyásolja negatívan a látogató élményét. Az egérmozgás és billentyűzet használat detektálásának lehetőségét teljesítményi okokból a végső alkalmazásból kivettem.

Szerver oldal

Mivel a szerveroldal kevésbé befolyásolja a kliensoldali felhasználói élményt az elmentendő adatok terhelését és relevanciáját a szerveroldali hardverigény és programozási technológia helyes megválasztása miatt tettem.

Tényező neve	Hozzáférés módja	Hardver terhelése	Relevancia
IP cím	környezeti változóként áll rendelkezésre, amennyiben a látogató proxy mögött van, nem a valós IP címét adja vissza	alacsony	magas
HTTP fejléc	szerveroldalon környezeti változóként rendelkezésre áll	alacsony	magas
operációs rendszer típusa és verziószáma	a HTTP fejléc User-Agent string-jéből API-n keresztül vagy egy ingyenesen hozzáférhető adatbázis[36] segítségével határozható meg	közepes	magas
böngésző típusa és verziószáma		közepes	magas
host	általában beépített parancs segítségével oldalható fel	közepes	magas
ISP	harmadik fél adatbázisából lekérdezhető (nem szükséges a lekérdezés időpontjában elvégezni)	magas	magas

[36] PHP UserAgent Master

csatlakozáshoz használt hálózat típusa	A szerveroldalon nem áll rendelkezésre	-	-
felhasználó böngészési előzménye	a kliensoldalon nem hozzáférhető, az előzőleg meglátogatott és a nyomkövető script által adatbázisba lementett előzmények a szerveroldalon lekérdezhetőek (nem szükséges a lekérdezés időpontjában elvégezni)	közepes	közepes
meglátogatott webhely bejárásának módja	a látogató böngészési előzményeiből kinyerhető adat (nem szükséges a lekérdezés időpontjában elvégezni)	magas	közepes
oldalletöltés ideje	kliens és szerveroldalon egyaránt detektálható	alacsony	magas
meglátogatott URL	a kliens és szerveroldalon egyaránt környezeti változóként elérhető	alacsony	magas
generált eszközböngésző azonosító	a szerver által generált azonosító, amelyet vissza kell juttatni a kliens számára, majd ott elmenteni sütibe és localStorage-ba	közepes	magas
közösségi hálózatok által hozzáférhető adatok	felhasználó hozzájárulása szükséges, harmadik féltől, célszerű a szerveroldalon letölteni, mivel a sok adat letöltése a kliens által ronthatja a böngészési élményt, valamint, a felhasználó megszakíthatja a műveletet,	magas	magas

	így értékes információkat veszíthet el a nyomkövető alkalmazás		
harmadik féltől hozzáférhető adat	szerveroldalon letölthető harmadik féltől	közepes - magas	közepes – magas

A közösségi hálózatok által hozzáférhető adatok hardverterhelése igen magas, hiszen a lekérdezés több 10 másodpercig is eltarthat az áttöltendő adatok mennyiségének függvényében, emiatt ezt a részfeladatot egy külön szálon futtattam a lekérdezés blokkolása nélkül.

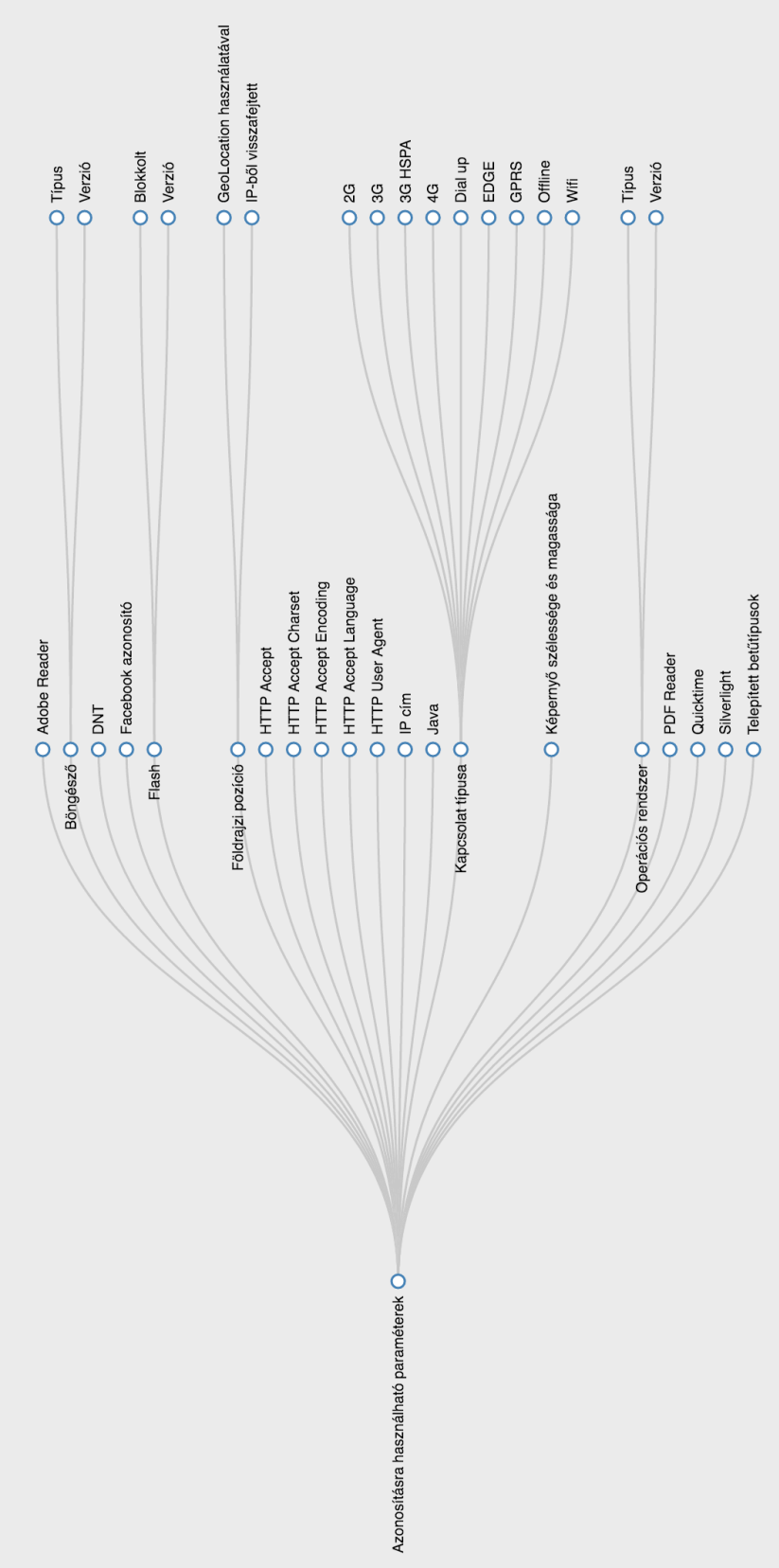
8.7.3. A lementett adatok listája

Az adatgyűjtő alkalmazás által lementett adatok teljes listája és az egyes értékek magyarázata az alábbi táblázatban olvasható.

MySQL adatbázis mező neve	A mezőben tárolt érték
userId	az eszközböngésző számára kiosztott azonosító
facebookId	a látogató Facebook azonosítója
facebookName	Facebook-on használt név
facebookFirstName	Facebook-on használt keresztnév
facebookLastName	Facebook-on használt vezetéknév
facebookGender	Facebook-on megadott nem
facebookUsername	Facebook-on megadott felhasználónév
facebookBirthday	Facebook-on megadott születésnap
facebookHometown	Facebook-on megadott születési hely
facebookLocation	Facebook-on megadott jelenlegi helyzet
facebookFriends	Facebook-on lévő barátok listája
facebookComments	a látogató Facebook kommentjei
facebookLikes	a látogató Facebook like-jai
facebookFeed	a látogató Facebook fala
datetime	a record elmentésének dátuma
IP	az internetezéshez használt eszközböngésző IP címe
host	az IP címből feloldott kiszolgáló neve
screenWidth	a látogató eszközböngészőjének szélessége

screenHeight	a látogató eszközböngészőjének magassága
httpUserAgent	a látogató eszközböngészőjének User-Agent string-je
httpAccept	a látogató eszközböngésző által támogatott állományok MIME formátuma
httpAcceptEncoding	a látogató eszközböngészője által támogatott adatátvitel kódolási típusok
httpAcceptLanguage	a látogató eszközböngészője által támogatott nyelvek és súlyozásuk
httpAcceptCharset	a látogató eszközböngészője által támogatott karakterkódolások és súlyozásuk
httpDNT	a látogató eszközböngészőjének Do-Not-Track értéke
applicationCache **history** **audio** **video** **indexedDB** **localStorage** **sessionStorage** **webSockets** **webSQLDatabase** **webWorkers** **geoLocation** **touch** **webGL**	HTML 5 képességek támogatása
Flash **Silverlight** **QuickTime** **Java** **PDFReader** **AdobeReader**	az eszközböngészőre telepített harmadik fél által készített kiegészítők és verziószámuk

referrer	a látogató által megtekintett e-learning oldal URL-je
connectionType	a csatlakozáshoz használt kapcsolat típusa, csak mobil eszközök esetében használt
isFlashBlocked	a látogató eszközböngészőjére Flash blokkoló telepítve van
isGeo	ha az értéke 1, akkor a postitionLatitude és a positionLongitude mezők értéke a GeoLocation API-ból érkezett ha az értéke 0, akkor a postitionLatitude és a positionLongitude mezők értéke IP címből visszafejtett hozzávetőleges pozíció
positionLatitude	a látogató eszközböngészőjének földrajzi pozíciójának szélességi foka
positionLongitude	a látogató eszközböngészőjének földrajzi pozíciójának hosszúsági foka
fonts	a böngészéshez használt eszközre telepített betűtípusok listája
fontsCounter	a böngészéshez használt eszközre telepített betűtípusok száma
browserFamily	a böngészéshez használt böngésző típusa
osFamily	a böngészéshez használt eszközböngésző operációs rendszere
osVersion	a böngészéshez használt eszközböngésző operációs rendszerének verziója
device	a böngészéshez használt eszköz



30. ábra: Azonosításra használható eszközböngészőből kinyerhető paraméter és felvehető értékei (saját szerkesztés)

8.7.4. A lementett adatok statisztikai jellemzői

Az adatgyűjtési időszak leteltét követően kerülhetett sor az adatbázisba elmentett paraméterek statisztikai vizsgálatára. Az alábbi táblázatban összefoglaltam a lementett paraméterek néhány alapvető statisztikáját: a mintában található kitöltött értékek száma és aránya, a különböző értékek számossága és aránya, valamint a különböző értékek eszközböngészőkhöz viszonyított aránya.

Paraméter neve	Kitöltött érték	Különböző értékek	Különböző értékek a kiosztott eszközböngészők arányában
Kiosztott egyedi azonosító (userID)	647 242 (100%)	32 529 (5,02%)	-
Facebook azonosító (userID)	176 (0,02%)	139 (0,02%)	0,42%
IP cím	647 242 (100%)	31 781(4,91%)	**97,70%**
Képernyő magassága és szélessége	647 242 (100%)	1 490 (0,23%)	4,58%
HTTP user agent	647 242 (100%)	2 173 (0,3%)	**6,68%**
HTTP accept	647 242 (100%)	9 (0,00%)	0,02%
HTTP accept encoding	647 242 (100%)	13 (0,00%)	0,03%
HTTP accept language	647 242 (100%)	9 (0,00%)	0,02%
HTTP accept charset	647 242 (100%)	22 (0,00%)	0,07%
DNT	647 242 (100%)	2 (0,00%)	0,01%
Böngésző típusa	647 242 (100%)	7 (0,00%)	0,02%
Böngésző típusa és verziószáma	647 242 (100%)	206 (0,03%)	0,63%
Operációs rendszer típusa	647 242 (100%)	8 (0,00%)	0,02%
HTML 5 képességek (application cache,	647 242 (100%)	2 (0,00%)	0,01%

history, audio, video, indexedDB, localStorage, sessionStorage, webSocket, webSQL, web workers, geoLocation, érintőképernyő, webGL)			
Flash verziója	624 459 (96,47%)	104 (0,02%)	0,32%
Silverlight verziója	502 649 (77,66%)	25 (0,00%)	0,08%
Quicktime verziója	250 959 (38,77%)	38 (0,01%)	0,12%
Java verziója	348 693 (53,87%)	50 (0,01%)	0,15%
PDF Reader verziója	534 152 (82,52%)	1 (0,00%)	0,01%
Adobe Reader verziója	334 969 (51,75%)	89 (0,01%)	0,27%
URL (referrer)	647 242 (100%)	17 353 (2,68%)	**53,35%**
Kapcsolat típusa	647 242 (100%)	4 (0,00%)	0,01%
Flash blokkoló használata	647 242 (100%)	2 (0,00%)	0,01%
Szélességi és hosszúsági fok	647 242 (100%)	2 (0,00%)	0,01%
Telepített betűtípusok	619 301 (95,68%)	14 895 (2,30%)	**45,78%**

33. táblázat: a lementett paraméterek statisztikai jellemzői

A táblázat utolsó oszlopában vastagon szedett betűkkel jelöltem azon paramétereket, amelyek esetében magas a különböző értékek aránya a mintában. (IP cím, HTTP user agent, URL (referrer) és telepített betűtípusok) A kiemelt értékek

közül az IP cím biztosan nem használható az eszközböngészők azonosítására, hiszen dinamikus IP-t használó eszközök esetében ez az érték gyakran változhat. A URL szintén a megtekintett tartalom függvényében változik, így marad a HTTP User-Agent, ami a böngésző és az operációs rendszer nevét és verziószámát tömöríti, valamint az eszközböngészőre telepített betűtípusok listája, amely böngészőfüggetlen, így akkor is alkalmas az eszköz azonosítására, ha a látogató böngészőt vált.

8.7.5. A feldolgozott adatok listája

Az adatok lementését követően következett az adatok előfeldolgozása és nominálissá való alakítása, mivel az Apriori algoritmus csak ilyen típusú inputváltozókat fogad el. Az alábbi listában foglaltam össze az adatbázisból az Apriori algoritmus input adatait és azok értelmezését:

- is_DNT_on: a látogató böngészőjének Do-Not-Track beállításának állapota
- has_flash: van-e Flash plugin telepítve a látogató böngészőjére
- is_flash_blocked: amennyiben van Flash plugin telepítve, az blokkolva van-e
- has_silverlight: van-e Silverlight plugin telepítve
- has_quicktime: van-e Quicktime plugin telepítve
- has_java: van-e Java plugin telepítve
- has_PDFReader: a látogató böngészőjének van-e PDF olvasási képessége
- has_AdobeReader: van-e Adobe Reader plugin telepítve a látogató böngészőjére
- connection_type: a mező értékei a HTML 5 Network Information API-jának értékei
    - 0 – Ismeretlen
    - 1 – Ethernet kapcsolat
    - 2 – WiFi
    - 3 – Mobiltelefonos 2G kapcsolat
    - 4 – Mobiltelefonos 3G kapcsolat
    - 5 – Mobiltelefonos 4G kapcsolat
    - 6 – Mobiltelefonos kapcsolat
    - 7 – Nincs kapcsolat
- browserFamily: a látogató böngészőjének típusa
- osFamily: a böngészéshez használt operációs rendszer típusa

- device: a böngészéshez használt eszköz típusa
- mobile_os_family: mobil operációs rendszer esetében annak neve
- font_counter: a látogató operációs rendszerére telepített betűtípusok száma
- screen_size: a képernyő felbontása (a képernyő szélessége alapján)
    - ha a képernyő szélessége <= 600, akkor svga
    - ha a képernyő szélessége > 600, de <= 1024, akkor xga
    - ha a képernyő szélessége > 1024, de <= 1280, akkor sxga
    - ha a képernyő szélessége > 1280, de <= 1400, akkor sxga_plus
    - ha a képernyő szélessége > 1400, de <= 1920, akkor hd
    - ha a képernyő szélessége > 1920, akkor 4k
- number_of_pageloads: a látogatóhoz tartozó oldalletöltések száma
- cardinality_of_sessions: a látogató által megkezdett munkamenetek száma
- geo_category: a böngészés helyszíne
    - 1: Európán kívül
    - 2: Európán belül
    - 4: Magyarország, Budapest
    - 5: Magyarország, megyeszékhely
    - 6: Magyarország, egyéb település
- cardinality_of_locations: különböző helyszínek számossága
- activity_dawn: 4:00-7:00 közötti oldalletöltések száma
- activity_forenoon: 7:00-10:00 közötti oldalletöltések száma
- activity_noon: 10:00-14:00 közötti oldalletöltések száma
- activity_afternoon: 14:00-18:00 közötti oldalletöltések száma
- activity_evening: 18:00-23:00 közötti oldalletöltések száma
- activity_night: 23:00-4:00 közötti oldalletöltések száma

Osztályváltozóként használt változók

Az alábbiakban az osztályváltozóként használt személyhez köthető változókat sorolom fel:

- course: a látogató által látogatott kurzus
- department: a látogató által látogatott kurzus tanszéke
- cardinality_of_likes: Facebook Like-ok számossága
- nem: 0 - férfi, 1 – nő

A myPersonality API által visszaadott 0 és 1 közötti értékek a vizsgált mintára vonatkozó percentilis értékek. [10]

- satisfaction_life: élettel való megelégedettség
- intelligence: intelligencia, amely az alábbi képlettel határozható meg:

$$IQ = 77{,}941 + 44{,}118 \times \text{intelligencia}$$

- s_yes: kapcsolatban
- r_christian: keresztény vallású
- b_agreeableness: Big5 barátságosság értéke
- b_conscientiousness: Big5 lelkiismeretesség értéke
- b_openness: Big5 nyitottság értéke
- b_neuroticism: Big5 neurotikusság értéke
- b_extraversion: Big5 extrovertáltság értéke
- p_uninvolved: nem érdekli a politika
- p_conservative: politikailag konzervatív

8.8. Idézett források

8.8.1. Tudományos szakirodalmi művek

Abramson, M., & Aha, D. W. (2013). User authentication from Web browsing behavior. *Florida Artificial Intelligence Research Society Conference* (old.: 6). St. Pete Beach: AAAI Press.

Agrawal, R., & Srikant, R. (1994). Fast Algorithms for Mining Association Rules. *VLDB '94 Proceedings of the 20th International Conference on Very Large Data Bases* (old.: 487-499). San Francisco: Morgan Kaufmann Publishers Inc.

András, N., & Péter, S. (2015). Empirical Analysis of Information Security Awareness in the Business and Public Sectors of Hungary. *Central and Eastern European e|Dem and e|Gov* (old.: 405-418). Wien: Leibniz Information Centre for Economics.

Andreas, P., & Marit, H. (2010. augusztus 10). *Privacy and Data Security, TU Dresden, Faculty of Computer Science*. Forrás: A terminology for talking about privacy by data minimization: Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management: http://dud.inf.tu-dresden.de/Anon_Terminology.shtml

Andrews, G., Gilbert, J., Repper, M., Roth, G., & Wear, J. (2011). *Internet Anonymity*. Forrás: Anonymity on the Internet - CS 181 Final Project, Spring 2011: https://sites.google.com/site/cs181anonymity/definition

Barabási, A. (2010). *Villanások - a jövő kiszámítható*. Budapest: Helikon Kiadó Kft.

Bodon, F. (2010. február 28). Adatbányászati algoritmusok. Budapest, Magyarország.

Brett, B. (2011). *The Psychology of Sharing: Why do people share online?* The New York Times.

Carlos, G.-U. A., & Neil, H. (2015. december). The Netflix Recommender System: Algorithms, Business Value, and Innovation. *ACM Transactions on Management Information Systems, 6*, 19.

Chris, H. J., Ashkan, S., Nathaniel, G., & Dietrich, W. J. (2012. january 1). Behavioral Advertising: The Offer You Can't Refuse. *Harvard Law & Policy Review vol. 6*, old.: 273-296.

Clarke, R. (1999). Internet Privacy Concerns Confirm the Case for Intervention. *Communications of ACM*, 60-67.

Cser, L., & Fajszi, B. (2004). *Üzleti tudás az adatok mélyén - Adatbányászat alkalmazói szemmel*. Budapest: Budapesti Műszaki és Gazdálkodástudományi Egyetem.

Cser, L., Nagyné Polyák, I., & Németh, Z. (2007). *Informatikai Alapok*. Debrecen: Debreceni Egyetem Agrár- és Műszaki Tudományok Centruma.

Danezis, G., Domingo-Ferrer, J., Hansen, M., Hoepman, J.-H., Le Métayer, D., Tirtea, R., & Schiffner, S. (2015. január 12). *European Union Agency for Network and Information Security*. Letöltés dátuma: 2017. május 14, forrás: Privacy and Data Protection by Design: https://www.enisa.europa.eu/publications/privacy-and-data-protection-by-design

Davenport, D. (2002. április). Anonymity on the Internet: Why the Price May Be Too High. *Communications of the ACM vol. 45, no. 4*, old.: 33-35.

Domokos, M. N. (2013). Az EU új adatvédelmi szabályozása – avagy „keep bangin' on the wall of Fortress Europe". *Jogi Fórum*, 1-46.

Eckersley, P. (2013, january 26). *Electronic Frontier Foundation - Defending your rights in the digital world*. Retrieved april 19, 2013, from A Primer on Information Theory and Privacy: https://www.eff.org/deeplinks/2010/01/primer-information-theory-and-privacy

ENISA. (2017). *European Union Agency for Network and Information Security*. Letöltés dátuma: 2017. május 14, forrás: About ENISA: https://www.enisa.europa.eu/about-enisa

Escobido, M., & Gillian, S. (2013). Can Personality Type be Predicted by Social Media Network Structures? *The Asian Conference on Psychology & the Behavioral Sciences*. Osaka: The International Academic Forum.

Európai Bizottság. (2015. július 11). *A személyes adatok védelme*. Forrás: Európai Bizottság honlapja: http://ec.europa.eu/justice/data-protection/index_hu.htm

France, B., & Robert, C. E. (2011). Privacy in the digital age: A review of information privacy research in information systems. *MISQ, volume 35, issue 4*, 1017-1041.

Golbeck, J., Robles, C., & Turner, K. (2011). Predicting personality with social media. *CHI'11 Extended Abstracts on Human Factors in Computing Systems* (old.: 253-262). New York: ACM New York.

Haig, Z., Kovács, L., Ványa, L., & Vass, S. (2014). *Elektronikai hadviselés*. Budapest: Nemzeti Közszolgálati Egyetem.

Hunyadi, L., & Vita, L. (2006). *Statisztika közgazdászoknak*. Budapest: Központi Statisztikai Hivatal.

Iváncsy, R., & Juhász, S. (2007). Analysis of Web User Identification. *International Journal of Computer Science;2007, Vol. 2 Issue 3*, 212.

Jeffrey, R. S. (2012). *Facebook: A Case Study of Strategic Leadership*. Zurich: Swiss Management Center - Transknowlogy Campus.

Jia-Ching, Y., Chu-Yu, C., & Vincent, T. S. (2012). Mining web navigation patterns with dynamic thresholds for navigation prediction. *IEEE Computer Society 2012* (old.: 614-619). Hangzhou: IEEE.

John, L., Manuel, B., & Luis, A. v. (2004. február). Telling Humans and Computers Apart Automatically. *Communications of the ACM*, 57-60. Letöltés dátuma: 2013. július 14, forrás: http://www.cs.cmu.edu/~biglou/captcha_cacm.pdf

Jurafsky, D., & Manning, C. (2012. april 4). Precision, Recall, and the F measure. *Stanford NLP*.

Kang, R., Brown, S., & Kiesler, S. (2013). Why do people seek anonymity on the internet?: informing policy and design. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2657-26666.

Kennedy, H. (2006). Beyond anonymity, or future directions for internet identity research. *New Media & Society, Vol 8, Issue 6*, 859-876.

Kiss, A. (2015. február 23). *Az adatokhoz, adatbázisokhoz kapcsolódó jogi szabályozás 1*. (A. Kiss, Előadó) Budapest.

Kosinksi, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *PNAS*, 5802-5805.

Kosinksi, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. In U. o. Kenneth Wachter (Szerk.), *Proceedings of the National Academy of Sciences of the United States of America. 110*, old.: 5802–5805. Berkeley: PNAS.

Kosinski, M., Bachrach, Y., Kohli, P., Stillwell, D., & Graepel, T. (2013. october 19). Manifestations of user personality in website choice and behaviour on online social networks. *Machine Learning, June 2014, Volume 95*, old.: 357-380.

Kosinski, M., Las Casas, D., Paulo Pesce, J., Quercia, D., Stillwell, D., Almeida, V., & Crowcroft, J. (2012). Facebook and Privacy: The Balancing Act of Personality, Gender, and Relationship Currency. *Sixth International AAAI Conference on Weblogs and Social Media*. Dublin: ICWSM.

Kovács, E. (2014). *Többváltozós adatelemzés*. Budapest: Typotex.

Kovács, L. (2013). Cyberterrorizmus: valós vagy túldimenzionált veszély? *Magyar rendészet XIII. (különszám)*, old.: 85-93.

Mano, K. (2011. november). Mathematical Duality between Anonymity and Privacy and Its Application to Law. *NTT Technical Review Vol. 9, No. 11*, old.: 1-7.

Marx, G. T. (1999). What's in a Name? Some Reflections on the Sociology of Anonymity. *The Information Society*, 99-112.

Mirnics, Z. (2006). *A személyiség építőkövei*. Budapest: Bölcsész Konzorcium.

Nan, Z., Aaron, P., & Haining, W. (dátum nélk.). *An Efficient User Verification System via Mouse*.

Nemeslaki, A., & Pocsarovszky, K. (2011. szeptember 18). Web crawler research methodology. *Web Crawler Research Methodology*, (old.: 1-17). Budapest.

Nemeslaki, A., & Sasvári, P. (2015). Empirical Analysis of Information Security Awareness in the Business and Public Sectors of Hungary. *Central and Eastern European e|Dem and e|Gov Days 2015* (old.: 405-418). Bécs: Druckerei Riegelnik.

Nemeslaki, A., Kis, G., Duma, L., & Szántai, T. (2004). *e-Business: Üzleti modellek*. Budapest: ADECOM Kommunikációs Szolgáltató Rt.

Peter, O., David, G., David, L., Warren, F., & Jonathan, N. B. (2005). Continuous Identity Verification. *Jur*, 20-24.

Poria, S., Gelbukh, A., Agarwal, B., Cambria, E., & Howard, N. (2013). Common Sense Knowledge Based Personality Recognition from Text. *Advances in Soft Computing and Its Applications* (old.: 484-496). Mexico City: Springer Berlin Heidelberg.

Racskó, P. (2012). A számítási felhő az Európai Unió Egén. *Vezetéstudomány*, old.: 1-16.

Regina, D. E. (2012. november 13). *Y és Z generáció mint a jövő munkavállalói*. Forrás: Kormányhivatal: http://www.kormanyhivatal.hu/download/2/18/60000/Y%20%C3%A9s%20Z%20gener%C3%A1ci%C3%B3%20mint%20a%20j%C3%B6v%C5%91%20munkav%C3%A1llal%C3%B3i.pdf

Rétallér, O., & Balogh, Z. (2015. december). Specialities of Psychological Traits of Citizens of Corvinus University of Budapest. *Hadmérnök*, 15.

Rigby, K. (1995). Anonymity on the Internet Must be Protected. *Paper for MIT*.

Robyn, R. L., Krishen, A. S., & Kachroo, P. (2014). Understanding the Components of Information Privacy Threats for Location-Based Services. *Journal of Information Systems*, 227-242.

Rössler, B. (2004). *The Value of Privacy*. Amsterdam: John Wiley & Sons.

Sander, T., Majláth, M., Sloka, B., & Lee Teh, P. (2015). User preference and channels use in the employment seeking process. *Management, Enterprise and Benchmarking in the 21st century II*. (old.: 239-248). Budapest: Óbuda University.

Shababi, C., Zarkesh, M. A., Adibi, J., & Shah, V. (1997). Knowledge discovery from users web page navigation. *26th IEEE International Conference on research in Data Engineering*, (old.: 20-29).

Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*, 379-423.

Shivani, H. (2004). *Authentication by Mouse Movements*.

Stillwell, D. J., & Kosinki, M. (2012). *myPersonality project: Example of successful utilization of online social networks for large-scale social research*. Cambridge, University of Cambridge, UK: The Psychometrics Centre.

Stillwell, D., Kosinki, M., Rust, J., & Wang, N. (2012. february 3). Can Well-Being be Measured Using Facebook Status Updates? Validation of Facebook's Gross National Happiness Index. *Social Indicators Research vol 115, issue 1*, old.: 483-491.

Szabó, A. (2010). *Random Forests - Véletlen erdők*. Letöltés dátuma: 2017. január 8, forrás: Adatbányászat és Keresés Csoport: https://dms.sztaki.hu/sites/dms.sztaki.hu/files/file/2011/randomforests.pdf

Személyes adatok feldolgozása vonatkozásában az egyének védelméről és az ilyen adatok szabad áramlásáról, 95/46/EK (Az Európai Parlament és a Tanács 1995. október 24).

Szommer, K., Balogh, Z., & Racskó, P. (2014). Az online világban hagyott virtuális lábnyomokban rejlő információ és azok veszélyei. *Vezetéstudomány*.

Voulodimos , A. S., & Patrikakis , C. Z. (2009. december). Quantifying privacy in terms of entropy for context aware services. *Identity in the Information Society, 2*(2), 155-169.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining - 3rd edition*. Burlington: Morgan Kaufmann.

Wolf, R., Nemeslaki, A., & Sasvári, P. (2015). Old Monarchy in the New Cyberspace: Empirical Examination of Information Security Awareness among Austrian and Hungarian Enterprises. *Academic and Applied Research in Military and Public Management Science*, 63-78.

Xu, H., Dinev, T., & Smith, J. H. (2011). Information privacy research: An interdisciplinary review. *MISQ, Volume 35, Issue 4*, 989-1015.

Youyou, W., Kosinki, M., & Stillwell, D. (2015. január 27). Computer-based personality judgments are more accurate than those made by humans. *PNAS*, old.: 1036-1040.

8.8.2. Hírek és cikkek

Adobe. (2011. július). *Adobe*. Letöltés dátuma: 2013. augusztus 25, forrás: Macromedia Flash and Shockwave Players: Macromedia Flash and Shockwave Players

Adobe. (2011. július). *Adobe*. Letöltés dátuma: 2013. augusztus 10, forrás: Adobe Flash Platform runtimes / Statistics : PC penetration: http://www.adobe.com/mena_en/products/flashplatformruntimes/statistics.html

Alessandro, A. (2013. június). *TED*. Forrás: What will a future without secrets look like?: http://www.ted.com/talks/alessandro_acquisti_why_privacy_matters

Alexander, D. (2012. augusztus 24). *Scatmania - The adventures and thoughts of "Scatman" Dan Q*. Letöltés dátuma: 2013. július 27, forrás: Visitor Tracking Without Cookies (or How To Abuse HTTP 301s): http://www.scatmania.org/2012/04/24/visitor-tracking-without-cookies/

Ali, R. (2010. július 13). *Hasseg.org*. Forrás: Getting a List of Installed Fonts with Flash and Javascript: http://hasseg.org/blog/post/526/getting-a-list-of-installed-fonts-with-flash-and-javascript/

All About Cookies. (2011. május 26). *Cookies - Free Cookie Resources*. Letöltés dátuma: 2013. augusztus 11, forrás: Welcome To All About Cookies.org: http://www.allaboutcookies.org/

Andersen, A. (2008. szeptember 3). *History of the browser user-agent string*. Forrás: WebAIM - Web Accessibility in Mind: http://webaim.org/blog/user-agent-string-history

Berners-Lee, T. (2014. március). *TED*. Forrás: A Magna Carta for the web: http://www.ted.com/talks/tim_berners_lee_a_magna_carta_for_the_web?language=en#t-47545

Bhavin, R. (2014. május 6). *PHP And MySQL Training In Vadodara With Live Project*. Forrás: Joomla Training In Vadodara: http://dotnettrainingvadodara.blogspot.hu/

Brafton Editorial. (2013. május 28). *Brafton - Fuel your brand*. Letöltés dátuma: 2013. augusztus 25, forrás: Teens' data sharing attitudes make audience targeting easier: http://www.brafton.com/news/teens-data-sharing-attitudes-make-audience-targeting-easier

Built With. (2013. augusztus). *Built With*. Letöltés dátuma: 2013. augusztus 31, forrás: OpenID Usage Statistics: http://trends.builtwith.com/docinfo/OpenID

Camilla, T. (2015. március 27). *Facebook completes first drone flight above UK, Mark Zuckerberg confirms*. Forrás: The Telegraph: http://www.telegraph.co.uk/news/media/11499142/Facebook-completes-first-drone-flight-above-UK-Mark-Zuckerberg-confirms.html

Chaffey, D. (2015. április 21). *Display advertising clickthrough rates*. Forrás: Smart Insights: http://www.smartinsights.com/internet-advertising/internet-advertising-analytics/display-advertising-clickthrough-rates/

Chen, A. (2011. július 27). *Gawker*. Letöltés dátuma: 2011. november 20, forrás: Mark Zuckerberg's Sister: 'I Think Anonymity on the Internet Has to Go Away': http://gawker.com/5825343/mark-zuckerbergs-sister-i-think-anonymity-on-the-internet-has-to-go-away

Christina, F. (2013. december 27). *'Facebook is simply not cool anymore' to teens, study finds*. Letöltés dátuma: 2014. március 30, forrás: VentureBeat: http://venturebeat.com/2013/12/27/facebook-is-simply-not-cool-anymore-to-teens-study-finds/

Cisco. (2014. június 10). *Cisco*. Forrás: The Zettabyte Era—Trends and Analysis: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/VNI_Hyperconnectivity_WP.html

D. Blondel, V., A. Hidalgo, C., Verleysen, M., & de Montjoye, Y.-A. (2013, march 25). *Scientific Reports*. Retrieved may 4, 2013, from Unique in the Crowd: The privacy bounds of human mobility: http://www.nature.com/srep/2013/130325/srep01376/full/srep01376.html

Dan, C. (2004. szeptember 1). *Hypertext Transfer Protocol -- HTTP/1.1 - Method definitions*. Letöltés dátuma: 2013. július 14, forrás: World Wide Web Consortium (W3C): http://www.w3.org/Protocols/rfc2616/rfc2616-sec9.html

eNet. (2013). *Jelentés az internetgazdaságról*. Budapest.

Facebook Developers. (2013. július 12). *Facebook Developers*. Letöltés dátuma: 2013. július 15, forrás: Permissions: https://developers.facebook.com/docs/facebook-login/permissions/

Falkvinge, R. (2013. october 2). *Privacy Online News - Protect Your Privacy*. Forrás: How Does Privacy Differ From Anonymity, And Why Are Both Important?: https://www.privateinternetaccess.com/blog/2013/10/how-does-privacy-differ-from-anonymity-and-why-are-both-important/

Gary, K. (2012. február). *TED*. Forrás: Nyomon követni a nyomkövetőket: http://www.ted.com/talks/gary_kovacs_tracking_the_trackers?language=hu#t-368530

Halász, B. (2014. november 24). *Az EU új adatvédelmi rendelete - Jön! Jön. Jön?* Forrás: Twobirds ideas Hungary - A Bird & Bird blog: http://twobirdsideas.hu/2014/11/24/az-eu-uj-adatvedelmi-rendelete-jon-jon-jon/

Halász, B. (2016. május 4). *2018. május 25-étől kell alkalmazni az EU új adatvédelmi rendeletét!* Letöltés dátuma: 2017. március 26, forrás: Twobirds ideas Hungary - A Bird & Bird blog: https://twobirdsideas.hu/2016/05/04/2018-majus-25-etol-kell-alkalmazni-az-eu-uj-adatvedelmi-rendeletet/

Hallen, E. (2012. február 4). *Quora*. Forrás: How many friends does a Facebook user have on average, and what is the distribution of friends numbers?: https://www.quora.com/How-many-friends-does-a-Facebook-user-have-on-average-and-what-is-the-distribution-of-friends-numbers

Healy, B. (2015. június 15). *People voted on a Magna Carta for the Internet and here's what they want*. Forrás: Mashable: http://mashable.com/2015/06/15/magna-carta-digital-age-internet/#3EImqMRDfkqw

Heggestuen, J. (2013. december 15). *Business Insider*. Forrás: One In Every 5 People In The World Own A Smartphone, One In Every 17 Own A Tablet: http://www.businessinsider.com/smartphone-and-tablet-penetration-2013-10

Hobson, A. (2013. február 15). *Facebook tracks data from users who have logged out*. Letöltés dátuma: 2013. július 7, forrás: The Daily Caller: http://dailycaller.com/2013/02/15/facebook-tracks-data-from-users-who-have-logged-out/

JanRain. (2013). *JanRain - Registration is hard. We make it easy*. Letöltés dátuma: 2013. augusztus 31, forrás: JanRain Engage - Social Login & Share: http://janrain.com/

Kozma, Z. (2015. augusztus 5). *Még idén elfogadhatják az EU új adatvédelmi rendeletét*. Forrás: Advocatus - A DLA Piper jogi blogja: http://blogs.dlapiper.com/advocatus/?p=1692

Lohr, S. (2009. szeptember 21). *A $1 Million Research Bargain for Netflix, and Maybe a Model for Others*. Letöltés dátuma: 2017. február 19, forrás: The New York Times: http://www.nytimes.com/2009/09/22/technology/internet/22netflix.html?_r=0

Mathews, L. (2013. március 5). *Geek.com*. Letöltés dátuma: 2013. augusztus 17, forrás: The state of Do Not Track in web browsers: http://www.geek.com/news/the-state-of-do-not-track-in-web-browsers-1541614/

Medic, M. (2014. április 7). *Creating and using CRUD stored procedures*. Forrás: SQLShack: http://www.sqlshack.com/creating-using-crud-stored-procedures/

Mozilla.org. (dátum nélk.). *Mozilla.org*. Letöltés dátuma: 2013. augusztus 17, forrás: Do Not Track: http://www.mozilla.org/en-US/dnt/

NetMarketShare. (2016. december). *NetMarketShare*. Letöltés dátuma: 2017. január 1, forrás: Desktop Operating System Market Share: http://www.netmarketshare.com/operating-system-market-share.aspx?qprid=10&qpcustomd=0

OpenID. (2015). *OpenID*. Forrás: OpenID: http://openid.net/

Owen, J. (2014. march 12). *The Independent*. Forrás: 25 years of the World Wide Web: Tim Berners-Lee explains how it all began: http://www.independent.co.uk/life-style/gadgets-and-tech/news/25-years-of-

the-world-wide-web-the-inventor-of-the-web-tim-bernerslee-explains-how-it-all-began-9185040.html

Pariser, E. (2011. március). *Beware online "filter bubbles"*. Forrás: TED.com: http://www.ted.com/talks/eli_pariser_beware_online_filter_bubbles

Parker, N. (2012. április 30). *NBNCo - Bringing broadband to lif*. Letöltés dátuma: 2013. augusztus 10, forrás: How many net-connected devices are in your home?: http://www.nbnco.com.au/blog/how-many-net-connected-gadgets-in-your-home.html

Rosenbush, S. (2013. október 30). *Facebook Tests Software to Track Your Cursor on Screen*. Forrás: Wall Street Journal: http://blogs.wsj.com/digits/2013/10/30/facebook-considers-vast-increase-in-data-collection/

Salesforce. (2013. június). *The Facebook Ads Benchmark Report*. Forrás: Salesforcesocial.com: http://www.salesforcemarketingcloud.com/wp-content/uploads/2013/06/The-Facebook-Ads-Benchmark-Report.pdf

Sam, O. (2015. február 13). *Stanford researchers develop method for tracking mobile devices using battery charge data*. Forrás: Apple Insider: http://appleinsider.com/articles/15/02/23/stanford-researchers-develop-method-for-tracking-mobile-devices-using-battery-charge-data

Shadmand, S. (2013. október 9). *Socialize*. Forrás: Everything you need to know about iOS's IDFA, IDV & Cookies Overview: http://blog.getsocialize.com/2013/everything-you-need-to-know-about-ioss-idfa-idv-cookies-overview

Shall, B. (2017). *Useragent String Lookup*. Forrás: UseragentAPI: https://useragentapi.com/

Smart, J. (2013. december 17). *James Smart személyes Twitter oldala*. Letöltés dátuma: 2017. április 29, forrás: Twitter: https://twitter.com/jamessmat/status/412920722407686145

Smith, C. (2015. november 13). *By the Numbers: 200+ Amazing Facebook Statistics (November 2015)*. Forrás: DMR - Digital Marketing Stats/Strategy/Gadgets: http://expandedramblings.com/index.php/by-the-numbers-17-amazing-facebook-stats/

Stack Exchange. (2011. november 9). *Stack Exchange*. Forrás: Can web sites detect whether you are using private browsing mode?:

http://security.stackexchange.com/questions/9037/can-web-sites-detect-whether-you-are-using-private-browsing-mode

Stanclift, M. (2008. szeptember 10). *A quick look at browser engines*. Forrás: Jinson Joseph Elayidom – My Musings: https://getjins.wordpress.com/2008/09/10/a-quick-look-at-browser-engines-trident-gecko-webkit-presto/

Statista - The Statistics Portal. (2015). *Statista - The Statistics Portal*. Forrás: Average number of Facebook friends of users in the United States as of February 2014, by age group: http://www.statista.com/statistics/232499/americans-who-use-social-networking-sites-several-times-per-day/

Statistic Brain. (2014. január 1). *Statistic Brain*. Letöltés dátuma: 2014. március 18, forrás: Facebook statistics: http://www.statisticbrain.com/facebook-statistics/

Tapad. (2013). *Unify Life Across Devices*. Forrás: Tapad: http://www.tapad.com

The Onion Router. (2015. december 10). *The Onion Router*. Forrás: Tor: Overview: https://www.torproject.org/about/overview.html.en

The Psychometrics Centre. (2013). *Apply Magic Sauce - Prediction API*. Forrás: University of Cambridge - The Psychometrics Centre: http://applymagicsauce.com/you.html

The Stanford Review. (2014. május 2). *The Stanford Review - Stanford's Independent Newspaper*. Forrás: Starting today, Yahoo will not honor Do Not Track settings: http://stanfordreview.org/article/starting-today-yahoo-will-not-honor-do-not-track-settings/

Twitter Bootstrap. (2014. november 12). *Twitter Bootstrap*. Forrás: CSS: http://getbootstrap.com/css/#grid-media-queries

Vincenzo, C. (2016. január). *VincosBlog*. Letöltés dátuma: 2016. április 10, forrás: World Map of Social Networks: http://vincos.it/world-map-of-social-networks/

W3Schools. (2016. november). *Browser Display Statistics*. Letöltés dátuma: 2017. január 1, forrás: W3Schools: http://www.w3schools.com/browsers/browsers_display.asp

Websiteoptimization.com. (2009. február 19). *Anatomy of an HTTP request and correlation to Pagetest legend*. Forrás: Websiteoptimization.com: http://www.websiteoptimization.com/secrets/metrics/10-21-http-request.html

Yahoo Privacy Team. (2014. április 30). *Yahoo's Default = A Personalized Experience*. Forrás: Yahoo! Global Public Policy:

http://yahoopolicy.tumblr.com/post/84363620568/yahoos-default-a-personalized-experience

Zawadziński, M. (2013, april 26). *Ad Technology and Analytics*. Retrieved may 12, 2013, from Alternatives to cookie tracking: http://zawadzinski.com/2013/04/26/alternatives-to-cookie-tracking/

Zephora - Digital Marketing. (2017. március 7). *The Top 20 Valuable Facebook Statistics – Updated March 2017*. Forrás: Zephora - Digital Marketing: https://zephoria.com/top-15-valuable-facebook-statistics/

Zuckerberg, M. (2015. március 25). *2015 Opening keynote*. Forrás: Facebook Developer Conference: https://fbf8.com/