A Kongresszusi Könyvtár Twitter archívuma

Előzmények

2010-ben a *Library of Congress* (*LOC*) bejelentette, hogy megállapodott a legnépszerűbb mikroblog-szolgáltatóval, a Twitterrel az összes nyilvános tweet archiválásáról. A legfeljebb 140 karakterből álló rövid webes üzeneteket 2006 márciusáig (ekkor indult a Twitter) visszamenőleg is megkapja a könyvtár, előremenetben pedig folyamatosan átveszi az újakat. A szerződés megkötésekor az archív anyag mérete nagyjából 170 milliárd tétel volt, a napi gyarapodás pedig meghaladta az 50 milliót – és ez az átlagérték 2014-re megtízszereződött.

A Twitter az internet nagy nyilvános fóruma, ahol gyorsan és egyszerűen lehet információkat megosztani. Bár egyesek szerint főleg csak „céltalan fecsegés" zajlik rajta, mások viszont azt hangsúlyozzák, hogy milyen fontos szerepe van a hírszolgáltatás, a politikai kampányok, a vészhelyzetek kezelése, a szervezeten belüli és az ügyfelekkel való kommunikáció terén, vagy akár a nagy sport- és médiaesemények együttes átélésekor. A társadalomkutatók hamar felismerték, hogy a Twitter 284 milliós aktív felhasználói tábora által generált üzenetfolyam értékes adatforrás lehet például az erős és gyenge közösségi kapcsolatok, vagy a divatok és trendek időbeli változásának vizsgálatához. A Kongresszusi Könyvtár tervezett archívuma nagyszerű lehetőség lenne arra, hogy ez az érdekes digitális tartalom gondozott formában, hosszú távon is megmaradjon és elemezhető legyen a szakemberek számára.

Mivel a felhasználóknak csak kevesebb mint 10 százaléka korlátozza az üzenetei és adatai nyilvánosságát, a Twitter eddig is biztosított hozzáférést a kutatóknak adatbázisa publikus részéhez. Egy közelmúltbeli szakirodalom-kutatás szerint legalább 380 publikáció született már a Twitter-adatok elemzéséből a legkülönbözőbb diszciplínákban: például informatika, kommunikációtudomány, közgazdaságtan, társadalom- és viselkedéstudományok, nyelvészet és egyéb humán szakterületek. Az elemzésekhez felhasznált tweetek száma a néhány tucatnyitól a milliárdos nagyságrendig terjedt, és az adatokat vagy közvetlenül a Twitter webes felületéről, vagy az alkalmazásprogramozási interfészen (API-n) keresztül gyűjtötték be a cikkek szerzői. 2011 elején a szolgáltató jelentős változásokat vezetett be az API-ban és a használati feltételekben, erősen korlátozta az adatokhoz való hozzáférést, így gyakorlatilag ellehetetlenítette azokat a külső szolgáltatásokat (pl. a TwapperKeeper-t és a 140kit-et), amelyeket a kutatók előszeretettel használtak a Twitteren folyó élet monitorozására. A bevezetett korlátozások egyrészt az API-n keresztül való adatkérések gyakoriságát, másrészt a lekérhető tweetek számát érintették – utóbbi 1 és 10 százalék közé lett lecsökkentve. Mivel a szűrőmechanizmus részleteit a Twitter nem hozta nyilvánosságra, ezért ez az adatszűrés bizonyos fajta kutatásoknál komoly módszertani bizonytalanságot jelent.

Az ingyenes API-szolgáltatás mellett létezik egy Twitter Firehose (tűzoltófecskendő) nevű csatorna is, amelyen a nyilvános tweetek 100 százaléka valós időben megjelenik, de ehhez csak néhány szervezetnek van hozzáférése, melyek pénzt kérnek a használatért. Ráadásul az ezen áramló adatmennyiség fogadása, szűrése és feldolgozása akkora számítástechnikai teljesítményt igényel, amit sok kutató nem tud megfizetni. 2014 elején a Twitter meghirdetett egy Twitter Data Grants nevű pályázatot, amelyre kutatási tervekkel lehetett jelentkezni. Ám a több mint 1300 pályázóból mindössze 6 nyert ingyenes hozzáférést a teljes Twitter adatbázishoz. Ilyen esélyek mellett az a bejelentés, hogy a Library of Congress megkapja az egész archívumot, igazi örömhír volt a Twitter-elemzéssel foglalkozóknak, mert felcsillant a remény, hogy elhárulnak az eddigi akadályok az adatbázis használata elől.

Az amerikai Kongresszusi Könyvtár nemcsak hagyományos értelemben a legnagyobb könyvtár a világon (több mint 36 millió könyv és egyéb nyom-

tatott kiadvány, valamint 121 milliós térkép, kézirat, fotó, film, hang- és videofelvétel, illetve egyéb különgyűjteményi anyag), hanem a Nemzeti Digitális Információs Infrastruktúra és Megőrzési Programja keretében jelentős mennyiségű digitális tartalmat is gyűjt, őriz és szolgáltat. 2000 óta működtet egy webarchívumot, melynek mérete 2014 márciusában 525 terabájt volt, a havi növekedése pedig kb. 5 terabájt. Ebbe a tevékenységbe illeszkedett be az a döntés, hogy a könyvtár felvállalja a sokmilliárdnyi tweet megőrzését is a jövő számára, melyek akkorra mai világunk politikai, kulturális és társadalmi eseményeinek, trendjeinek múltbéli lenyomatai lesznek.

A 2010. április 14-én aláírt kétoldalas ajándékozási szerződés előírásai szerint a könyvtár csak a nyilvános tweeteket kapja meg és csak 6 hónap késéssel jelentetheti meg az újakat. Továbbá nem tehet letölthetővé „jelentős mennyiséget" az archívumból, valamint csak „jóhiszemű" kutatóknak adhat hozzáférést. Néhány héttel később egy blogbejegyzésből az is kiderült, hogy az átadás előtt már törölt üzenetek nem lesznek archiválva, és az üzenetekbe belinkelt képeket vagy weboldalakat sem gyűjti be a könyvtár.

2013 elején a LOC kiadott egy tájékoztatást a projekt állásáról. Eszerint a 2006−2010 közötti, 170 milliárd tételes archív állomány mérete 133,2 terabájt lett, s megoldották a bejövő, „élő" üzenetfolyam biztonságos és fenntartható fogadását és őrzését is a Gnip nevű – a közösségi médiából származó adatok aggregálásával foglalkozó – vállalaton keresztül. Közölték azt is, hogy további magáncégek bevonására lesz szükség a technikai és infrastrukturális problémák kezeléséhez, melyek miatt egyelőre nem tudnak hozzáférést biztosítani az archívumhoz. A helyzet sajnos azóta sem változott: több mint öt évvel az első bejelentés után, 2015 nyarán továbbra is elérhetetlen a LOC Twitter archívuma.

Problémák

Mint minden könyvtári gyűjteményt, a Twitter archívumot is fel kell dolgozni, rendszerezni és valamilyen módon katalogizálni ahhoz, hogy a kutatók számára hasznos, értelmes módon hozzáférhetővé tehessék. Bár a Kongresszusi Könyvtárban megvan a szükséges tapasztalat a digitális tartalmak kezelésére, de a hatalmas és gyorsan növekvő Twitter üzenetfolyam eddig ismeretlen kihívást eredményezett. Nemcsak a rövid szövegeket – és esetleg a bennük levő linkeket – kell ugyanis feldolgozni és eltárolni, hanem azt a több mint 100-féle metaadatot is, amely minden egyes tweethez kapcsolódik. A technikai problémák mellett a legnagyobb nehézséget a hozzáférési módok és szabályok kidolgozása jelenti, hiszen itt számos etikai, valamint adat- és magánélet-védelmi aggály is felmerül.

A műszaki feladat nagyságát jól érzékelteti az a tény, hogy amikor a LOC 2012 végén megkapta a teljes 2006−2010 közötti Twitter anyagot, ezzel csaknem megduplázódott az akkori digitális gyűjteményének a tárhelyigénye. Ráadásul egyre nő a tweetek száma: öt év alatt 50 millióról 500 millióra emelkedett a napi átlag, és bizonyos események idején az ütem igencsak megugrik: 2013. augusztus 3-án a „Laputa – Az égi palota" anime tévés vetítése alatt a japán nézők 143 199-re tolták fel az egy másodpercen belül elküldött tweetek rekordját (az átlagos érték 5 700 tweet/sec). A Twitter persze folyamatosan fejleszti az infrastruktúráját és alakítja át úgy a rendszerét, hogy képes legyen lépést tartani a növekvő igényekkel. A Kongresszusi Könyvtárnak viszont nincsen erre elegendő forrása és munkaereje, így muszáj külső technológiai partnereket bevonnia a munkába.

Ha sikerül megoldani ennek a hatalmas adattömegnek a fogadását és feldolgozását, a következő probléma a hozzáférés módjának mikéntje. A 2013. januári közleményében a LOC nyilvánosságra hozta, hogy bár már mintegy 400 kérés érkezett, de még senkinek nem adtak hozzáférést az archívumhoz, mivel jelen állapotában egyetlen keresőkérdés lefuttatása 24 óráig tartana. Megfelelő hardver és szoftver hiányában csak egy „alapszintű" kereshetőséget céloztak meg, és 2014 közepére ígértek egy kísérleti verziót, de még ez sem készült el. Arról sincs információ, hogy az adatok feldolgozása és indexelése után mire lesz képes ez az egyszerű kereső. Az ideális az volna, ha az üzenetek szövegében való keresésen túl a találati halmaz szűrhető lenne metaadatok szerint is (pl. felhasználóra, hashtag-re, időszakra, IP címből valószínűsíthető földrajzi helyre).

A technikai nehézségek remélhetőleg idővel megoldódnak, viszont a jogi és erkölcsi kérdések esetében nincsenek mindenkit kielégítő megoldások. A könyvtárszakmai etika azt diktálja, hogy egyenlő és teljes hozzáférést kell adni mindenkinek az információforrásokhoz, de ez nem minden esetben valósul meg a gyakorlatban. A LOC is kapott már kritikát „cenzúrázás" miatt, legutóbb például azért,

mert blokkolta a Wikileaks webszerveréhez való hozzáférést az olvasótermi gépekről. Bár a Twitter maga is végez némi tartalomszűrést, de az ajándékozási szerződés a LOC számára is engedélyezi, hogy a „megőrzésre nem alkalmas" részeket eltávolítsa az archívumból. Nem tudni viszont, hogy a könyvtárban kik és milyen szempontok szerint fogják kiválogatni ezeket az „alkalmatlan" üzeneteket, és hogy az archívumnak ezt a megszűrését hogyan lehet összeegyeztetni a gondolat-, vélemény- és információszabadság általános elveivel.

Az archívum bejelentésének pillanatában megjelentek a személyes adatok, a magánszféra védelmével kapcsolatos aggályok is. Sok meglepett és frusztrált felhasználó ekkor szembesült először azzal, hogy a mulandónak és személyesnek szánt üzenetei megőrződnek, sőt kutathatók. A tweetek közel fele tartalmaz valamilyen személyes információt a feladójáról (pl. elérhetőség, tartózkodási hely, egészségi állapot). Ráadásul az üzenetek továbbküldésének (retweet) gyakorlata miatt néha zárt körnek szánt információk is kiszivárognak. Egy 80 millió Twitter fiókra kiterjedt kutatás közel 250 ezer olyan védett account-ot talált, melyeknek legalább egy nem publikus üzenetét valaki továbbosztotta egy nyilvános fiókból.

A privacy-sértéssel (titoktartással) kapcsolatos aggodalmakra a LOC szóvivője azt válaszolta, hogy az archivált tartalom már amúgy is nyilvánosan elérhető a weben és hogy a Twitter felhasználói a regisztráláskor elfogadták a szolgáltatási szerződésben levő feltételeket. Ezzel a szokásos „már amúgy is nyilvános" érveléssel csak az a baj, hogy azon a hamis kettősségen alapul, hogy egy információ vagy csak szigorúan nyilvános vagy csak szigorúan privát lehet, és figyelmen kívül hagyja a kontextust – jelen esetben azt, hogy eredetileg kiknek indította el az üzenetét valaki a Twitteren át és milyen elvárásai voltak annak sorsával kapcsolatban. Az embereknek azon túl, hogy teljesen zárttá teszik a Twitter fiókjukat, jelenleg nincs más eszközük arra, hogy az üzeneteiket ne őrizze meg az archívum. A Twitter rendszerében van lehetőség egy tweet törlésére (ilyenkor az nemcsak a felhasználó saját idővonaláról tűnik el, hanem a követőkéről, valamint a keresőből is, továbbá a változtatás nélkül továbbított retweet-ek is törlődnek). Viszont a LOC-nál már archivált anyagban ez a törlés természetesen nem történik meg, így a felhasználók elvesztik a kontrollt a korábbi online tevékenységük és a magánéleti információik felett. Egyelőre nem tudni, hogy a könyvtár bevezet-e majd valamilyen törlési, illetve kimaradási (opt-out) lehetőséget.

A Twitter archívum története már eddig is sok tanulsággal szolgált arra vonatkozóan, hogy mekkora és milyen sokfajta nehézséget jelent a modern digitális környezetünk könyvtári megőrzése. A technikai gondokon az üzleti szféra bevonásával remélhetőleg sikerült úrrá lenni. Az etikai és hozzáférési kérdésekben pedig a könyvtárak és archívumok szakmai szervezeteinek ajánlásai jelenthetnek útmutatót. Remélhetőleg nem kell még további öt évet várni arra, hogy mindezek a problémák elfogadható módon megoldódjanak, és a kutatók használatba vehessék ezt a hatalmas és különleges digitális gyűjteményt.

/ZIMMER, Michael: The Twitter Archive at the Library of Congress: Challenges for information practice and information policy. = First Monday, 20. évf. 7. sz. 2015.
http://firstmonday.org/ojs/index.php/fm/article/view/5619/4653/

(Drótos László)