DOI: 10.38146/BSZ.2023.7.6

A nyílt forrásból származó adatgyűjtés automatizálásának lehetőségei

Possibilities of Automated Open Source Information Gathering

Gulyás Attila
doktorandusz, ny. alezredes
Óbudai Egyetem,
Biztonságtudományi Doktori Iskola
gulyas.attila@phd.uni-obuda.hu

Absztrakt

Cél: A digitális forradalom korát éljük, amelyben a korszerű technológiáknak és ezek dinamikusan növekvő felhasználói táborának köszönhetően eddig soha nem látott mennyiségű, szabadon hozzáférhető, részben nemzetbiztonsági szempontból fontos adat keletkezik a kibertérben, amelyek összegyűjtése és elemzése messze túllépi az emberi teljesítőképesség határait. Ez az adatrobbanás indokolttá teszi az eddig nyílt forrásból történő adatgyűjtés újragondolását, beleértve az információgyűjtés és feldolgozás automatizálási lehetőségeinek vizsgálatát.
Módszertan: A tanulmány egy félautomatizált rendszer felépítését és működését vázolja fel, megvizsgálva a megvalósítás elméleti lehetőségeit és buktatóit. A tanulmány nem vizsgálja a rendszer létesítésének és üzemeltetésének jogi kereteit. Terjedelmi okok miatt a szerző a fő pontok felvillantására koncentrált, hiszen egy-egy téma részletes tárgyalása önmagában is meghaladná a cikk kereteit. A tanulmány elkészítésekor a szemléletformálást és gondolatébresztést tűzte ki elsődleges célként.
Megállapítások: A kutatás során a hazai és a külföldi idevágó szakirodalom tanulmányozásán túl, mivel a szerző évek óta tanulmányozza a dark webet, illetve a nyílt forrású információgyűjtés lehetőségeit, nagymértékben támaszkodott a saját kutatási eredményeire és tapasztalataira.
Érték: A tanulmányban bemutatott elméleti rendszer vizsgálata rámutatott arra, hogy az ilyen rendszerek a jelenlegi technikai fejlettség mellett – ideértve a mesterséges intelligencia felhasználásának korlátait – nem működtethetők emberi felügyelet nélkül, ezért csak a félautomata rendszer kialakítása tűnik megvalósítható opciónak. Az adatfeldolgozás ma még elképzelhetetlen az ezen a területen

jártas adatmérnökök nélkül. Ezek mellet a korlátok mellet is az ilyen rendszer rendkívüli mértékben felgyorsíthatja a nyílt forrásból történő adatok gyűjtését, hatékonyabbá téve a felderítő munkát. A gyakorlati életben természetesen nem feltétlenül kell minden elemet egyszerre megvalósítani vagy létrehozni, hiszen már egyes részegységek működtetése és integrálása a meglevő rendszerekhez is jelentősen felgyorsíthatja és hatékonyabbá teheti az adatgyűjtés folyamatát.

Kulcsszavak: OSINT, dark web, mesterséges intelligencia, adatvizualizáció

Abstract

Aim: The digital revolution, in which advanced technologies and their dynamically growing user base are generating unprecedented amounts of freely accessible data in cyberspace, some of which is of national security importance, and the collection and analysis of which is far beyond human capacity. This explosion of data justifies a rethinking of the way data has been collected from open sources, including the exploration of ways to automate the collection and processing of information.

Methodology: The study outlines the design and operation of a semi-automated system, examining the theoretical possibilities and pitfalls of its implementation. The study does not examine the legal framework for setting up and operating the system. For reasons of space, the author has concentrated on highlighting the main points, as a detailed discussion of a single topic would exceed the scope of this article. The primary objective of the study is to stimulate and stimulate reflection.

Findings: In addition to the research on the relevant literature in Hungary and abroad, as the author has been investigating the dark web and the possibilities of open source information gathering for many years, he has relied heavily on his own research and experience.

Value: The theoretical system analysis presented in the study showed that such systems cannot be operated without human supervision at the current state of the technological development, including the limitations of the use of artificial intelligence, and therefore only the design of a semi-automated system seems to be a feasible option. Today, data processing is still unthinkable without data engineers skilled in this field. In addition to these limitations, such a system can speed up the collection of data from open sources to an extraordinary extent, making Intelligence more efficient. In practice, of course, it is not necessary to implement or create all the elements at once, as the operation and integration of some of the components into existing systems can significantly speed up and improve the efficiency of the data collection process.

Keywords: OSINT, Dark Web, Artificial Intelligence, data visualisation

Bevezetés

Az utóbbi néhány évtizedben zajló digitális forradalom következtében új technológiák, platformok jelentek meg, amelyek a felhasználók mind szélesebb köre számára váltak elérhetővé. A felhasználószám exponenciális növekedése és az általuk előállított adatok korábban sohasem tapasztalt mennyisége hagyományos eszközökkel már kezelhetetlen. Ebben az emberi elmével felfoghatatlan nagyságú és összetettségű adathalmazban a manuális adatgyűjtés gyakran végzetes késedelemhez vezethet, ezért elengedhetetlen a nyílt forrású információgyűjtés újragondolása, miközben a rendvédelmi szerveknek, illetve a nemzetbiztonsági szolgálatoknak törvényben foglalt feladataik végrehajtása érdekében nagy hangsúlyt kell fektetniük a titkos úton szerzett és a nyílt forrásból származó információk szinergikus feldolgozására és hasznosítására.

Az olyan új technológiák, mint a mesterséges intelligencia (Artificial Intelligence, AI), amely eddig elképzelhetetlen távlatokat nyit meg nagymennyiségű adatok gyors feldolgozásában, párosulva az internet evolúciójával és a közösségi média használatának elterjedésével korábban nem látott mértékben változtatja meg mindennapi életünket és ezzel együtt biztonsági környezetünket, új kihívások elé állítva a nemzetbiztonsági és rendvédelmi szerveket.

Ezeknek az új technológiáknak köszönhetően ugyanakkor lehetővé válik a nyílt forrásból származó információk automatizált gyűjtése, majd az összegyűjtött adatok mesterséges intelligencia felhasználásával történő feldolgozása. Tekintve azonban a mesterséges intelligencia jelenlegi állapotát és gyengeségeit az ilyen rendszerek még igénylik az emberi felügyeletet, ezért jelenleg a félautomatizált működés tűnik reálisnak és kivitelezhetőnek. Egy ilyen rendszer emberi segítség mellett alkalmas lehet a szükséges információk automatikus összegyűjtésére, rendszerezésére, illetve korai előrejelző rendszerek részeként azonnali intézkedést igénylő eseményekre történő időbeni reagálás indítására, vagy a már bekövetkezett események következményeiről azonnali információk beszerzésére (Near Time Intelligence), esetleg ilyen események utólagos rekonstruálására.

A kutatás célja

Tanulmányomban egy félautomatizált rendszer felépítését és működését igyekeztem felvázolni, megvizsgálva a megvalósítás elméleti lehetőségeit és buktatóit.

A munkám során terjedelmi okokból kénytelen voltam a fő kérdések ismertetésére szorítkozni, hiszen egy-egy téma részletes kifejtése önmagában meghaladná e tanulmány keretit. A tanulmány céljaként elsősorban a gondolatébresztést és szemléletformálást tűztem ki magam elé.

Polgári célú és nemzetbiztonsági változat elhatárolása

A félautomata adatgyűjtő rendszerek napjainkban nem kizárólag nemzetbiztonsági vagy bűnügyi célból üzemeltethetők, hanem a civil szférában, különösen az üzleti világban is előnyhöz juthat az ehhez hasonló szolgáltatásra előfizető, vagy ilyen rendszert üzemeltető vállalkozás, hiszen ilyen módon nyomon követheti marketing tevékenységének alakulását, begyűjtheti a különböző fórumokon a termékeivel, illetve szolgáltatásaival kapcsolatos véleményeket, de a rendszer alkalmas lehet üzleti hírszerző tevékenységbe (Business Intelligence) történő integrálásra is.

A hasonló vonások mellett a nemzetbiztonsági (bűnügyi) változatnak lehetővé kell tennie a felhasználó – esetünkben az adott szolgálat – valódi kilétének elrejtését, hiszen a beazonosíthatóság nem feltétlenül jelenthet problémát egy kereskedelmi változat esetében, azonban komoly hátránnyal járhat a nemzetbiztonsági, illetve rendvédelmi munka esetében. A szolgálat kilétének elrejtése, illetve háttérben tartásának igénye megjelenik már az egyszerű híroldalak, weboldalak és a más regisztráció nélküli weboldal meglátogatásánál is. A magyarázat a látogatók nyomon követhetőségében, illetve, profilozhatóságában rejlik. Széles körben ismert, hogy a felhasználók tevékenysége az interneten nyomon követhető. A teljesség igénye nélkül ezt a célt szolgálja a Web Cookie, a Web Beacon, vagy az úgynevezett browser fingerprinting technológia. Ezek mellett a weboldalak üzemeltetői a keresőmotorok találati listáján való előkelő helyezés megszerzése, illetve az oldal hatékonyabb üzemeltetése érdekében a jól ismert keresőmotorok által erre a célra biztosított kódszeleteit (tracker) a látogató számára észrevehetetlenül beépítik a weboldalakba. Ezek segítségével a keresőóriások nyomon követhetik a felhasználókat a világhálón. Jól példázza ezt az a jelenség, amikor az interneten keresett vagy vásárolt termék reklámja még napokig megjelenik a teljesen más témájú webhelyek reklámmezőiben.

A legegyszerűbb weboldal üzemeltetője, illetve a tárhely szolgáltatója is tudhatja, hogy hogyan és honnan jutott a felhasználó a szóban forgó weboldalra (organikus keresés eredménye, navigálás vagy átirányítás stb.), milyen IP címről látogatták meg az adott weboldalt, milyen operációs rendszeren, milyen

böngészőt használt a felhasználó, a webhely melyik oldalán mennyi ideig tartózkodott és onnan hová navigált tovább. Azokat az adatokat, amelyek a felhasználó szándéka nélkül keletkeznek, járulékos adatoknak, az angol terminológiában „exhausted data”-nak nevezik.

A közösségi oldalak esetében a fent említett információk még számos mással is kiegészülnek. A felhasználóhoz adatmezők és rekordok százai tartozhatnak. Az oldalak üzemeltetői többet tudhatnak egy adott felhasználóról, mint akár a legközelebbi családtagjai. Ezeknek a vállalkozásoknak a reklámtevékenység hatékonyabbá tétele érdekében egyik legfőbb tevékenységük a profilozás, a felhasználók minél jobb megismerése. Ennek érdekében – többek között – naplózzák a felhasználók tevékenységét, keresési előzményeit, a megtekintéseket stb. Lehetőségeiknél fogva képesek lehetnek a különböző felhasználói viselkedés és tevékenység alapján olyan profilok létrehozására is, amelyek az állami szereplőket, vagy akár a szolgálatokat is azonosíthatják. A már említett technológiák segítségével képesek megállapítani, hogy mely felhasználói fiókokat látogatják ugyanarról a böngészőről. A felhasználói viselkedés további elemzésével az is megállapítható, hogy ugyanaz a személy használ-e egy másik felhasználói fiókot, váltogatva a profiljai között, vagy egy másik személy, például egy munkatárs vagy esetleg egy családtag.

Végül, de nem utolsó sorban nem kerülhetjük ki az internetes keresésekben rejlő kockázati tényezőket sem. A technológiai óriások, az úgynevezett GAFAM (Google, Amazon, Facebook, Apple, Microsoft), hogy csak néhányat említsünk, rögzítik a keresési kifejezéseket a felhasználó IP címével és más paramétereivel együtt, ezért ezek a szolgáltatók – meglátásom szerint – csak részben alkalmasak az OSINT feladatokkal kapcsolatos keresések végrehajtására. Léteznek alternatívák, amelyek pontosan a globális adatgyűjtés elleni mozgalmak indíttatására naplózásmentes szolgáltatásokat ajánlanak, mint például a DuckDuckGo keresőmotor, de léteznek természetesen más megoldások is.

A meta-keresők olyan speciális keresőmotorok, amelyek egyszerre több keresőmotor felé küldik el a kéréseiket, majd a beérkezett válaszokat egy algoritmussal rendezik. Ezen a módon egyszerre több keresőmotor érhető el. Magától érthető módon ebben az esetben is követelmény az anonimitást biztosító kapcsolat igénybevétele, hiszen ebben az esetben a meta-kereső üzemeltetője elől kell a felhasználó identitását elrejteni.

Az internetes kereséséknél egyébiránt anonim kapcsolaton keresztül csatlakozó, az adott feladatra (tematikára) célirányosan kifejlesztett meta-kereső alkalmazása lehet a legjobb megoldás.

A felsorolt nyomonkövetési technikák kockázati tényezőt jelentenek egy szolgálat számára, hiszen az ellenérdekelt titkosszolgálatok, a szélsőséges politikai

szervezetek, a terrorszervezetek, illetve a szervezett bűnözői kör a tevékenységüket mélyen konspirálva szervezik és végzik, ebből kifolyólag az információgyűjtő kilétének felfedése – az esetleges politikai vetületen túl – a későbbi feldolgozó, illetve nyomozati munkát hátráltatná, esetleg meg is akadályozná.

A nyílt forrásból származó adatok gyűjtése semmiképpen sem jelenti azt, hogy minden ilyen jellegű forrásból csak nyílt módon kellene adatot gyűjteni. Mindenképpen szakítani kell azzal a szemlélettel, amely szerint a nyílt forrásból, nyílt módon kell információt gyűjteni, hiszen a fent említett technikai szempontok miatt az információgyűjtés nemzetbiztonsági jellege semmiképpen sem hagyható figyelmen kívül.

Természetesen vannak olyan források, ahol ez megvalósítható, hiszen lehet televíziót nézni vagy rádiót hallgatni, nyomtatott sajtót olvasni, de az internettel kapcsolatos adatgyűjtésnél már megfontolandó valamely anonimizáló vagy leplező védelmi intézkedés bevezetése, illetve eljárások alkalmazása.

Mielőtt azonban tovább haladnánk elengedhetetlen néhány alapfogalom tisztázása.

Az OSINT fogalma

Mindenek előtt tisztáznunk kell mit is értünk OSINT vagy nyílt forrásból származó információgyűjtés alatt.

Az OSINT kifejezés minden olyan információforrásra és információra utal, amely valamely legálisan és nyíltan hozzáférhető forrásból származik, és valamely személlyel, szervezettel vagy jelenséggel kapcsolatos. Az alábbiakban két megközelítést választottam a számos elérhető értelmezés közül.

A Kaspersky kiberbiztonsági vállalat enciklopédiájának meghatározása szerint az OSINT nem más, mint az információgyűjtés egyik ága, amely emberekkel és szervezetekkel kapcsolatos, nyílt forrásból származó információkat gyűjt (Kaspersky IT ecyclopedia, n. d.) (URL6).

Egy másik megközelítés szerint az Amerikai Egyesült Államok hadseregének ATP 2-22-9 szabályzatának 1-1. pontja alapján az OSINT a felderítés egyik válfaja, amely a nyilvánosan elérhető forrásokból összegyűjtött és feldolgozott információkat időben juttatja el a hírigényt megfogalmazó megrendelőnek (US Army, 2012).

A nyílt forrásból származó információgyűjtés nem alacsonyabb értékű, vagy kevésbé fontos, mint a titkos vagy bizalmas forrásból származó információszerzés. Eklatáns példa a CIA Oszama Bin Laden vadász egysége, amely állításuk szerint a hajtóvadászat során az információk 90%-át nyílt forrásból szerezte. Az Europol továbbmegy, szerinte a terrorizmus elleni harcban az információk

95%-a nyílt forrásból származik (Hobbs, Moran & Salisbury, 2014).

Klasszikus OSINT forrásnak tekinthetők az útibeszámolók, sajtó- és nyomdaipari termékek, televízió és rádióműsorok, hírügynökségi jelentések, a nem minősített hivatali jelentések, beszámolók, jegyzőkönyvek, beszédek, előadások stb. Az összegyűjtött adatot angolul Open Source Data-nak (OSD) nevezik.

A digitális korszak és a közösségi média berobbanását követően ez ma már kiegészül a túlnyomórészt az interneten, beleértve az open weben, a deep és dark weben, és különösen a közösségi médiában elérhető információk végtelennek tűnő mennyiségével.

A következőkben álljon itt néhány adat a percenként keletkező adatok elképesztő mennyiségéről (Lackey, 2019).

2019-ben percenként keletkezett adatmennyiség:

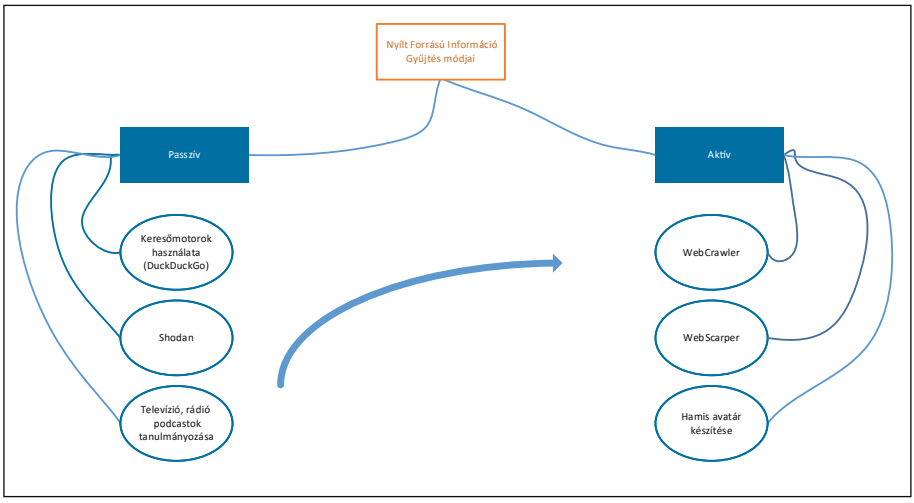
- több mint 474 000 tweet,
- 55 140 fotó poszt az Instagramon,
- 92 340 poszt a Tumblr-ön,
- 4 333 560 videót néztek meg a YouTube-on,
- 300 órányi videót töltöttek fel YouTube-ra,
- 510 000 Facebook-hozzászólás,
- 293 000 Facebook-státusz módosítása,
- 231 840 Skype-hívás,
- 4 497 420 keresés a Google keresőjében.

A fenti példa jól illusztrálja, hogy ilyen mennyiségű adatot már emberi erőforrásokkal átlátni, illetve kezelni lehetetlen.

Az OSINT gyűjtés módjai

Alapvetően kétfajta nyílt forrásból származó információgyűjtésről beszélhetünk. Az első esetben pusztán csak passzív módon összegyűjtjük az információt, például lementjük egy weboldal tartalmát, míg a második esetben az információ megszerzése érdekében valamilyen aktív cselekvés keretében regisztrálni kell, esetleg virtuális avatárt létrehozni. Ez utóbbi módszer nem összeegyeztethetetlen a nyílt forrásból származó információgyűjtéssel, mindaddig, amíg ez nem érinti a külső engedélyhez kötött információgyűjtés körébe tartozó információk megszerzését. Amint az 1. számú ábrán látható a passzív úton szerzett adatok felhasználhatók lehetnek az aktív módon végrehajtott információgyűjtés kiindulási alapjainak. A két módszer kategorikusan nem elválasztható, hanem egymással szorosan összefüggő, egymást kiegészítő rendszert alkotnak.

1. számú ábra
Az OSINT gyűjtés passzív és aktív módjai



Forrás: Az ábra a szerző saját szerkesztése.

Az OSINT felhasználási lehetőségei

A tanulmány további részében vizsgáljuk meg az OSINT néhány felhasználási területét. Napjaink gyorsan változó globális biztonsági környezetében a minél több forrásból származó, pontos, időbeni információ megszerzése elengedhetetlen feltétel. Ennek a követelménynek a teljesítéséhez az összadatforrású felderítésen belül egy célszerűen megtervezett, a törvényi előírásoknak megfelelő, félautomatizált, közel valós idejű (Near Time Intelligence) OSINT adatgyűjtő rendszer jelentősen hozzájárulhat. A rendszer felhasználásával gyűjtött adatok alapvetően az alábbi területeken jelentkező feladatok megoldásához használhatók fel.

1. Rossz szándékú befolyásolás
Ennek keretében az ellenérdekelt fél igyekszik az ország demokratikus, illetve gazdasági folyamataiba beavatkozni. Fogalmazhatunk úgy, hogy ide tartozik a politikai döntéshozatal és a közbeszéd befolyásolása. A közösségi média „meghekkelése", ami tulajdonképpen a különböző korlátozó algoritmusok kijátszása, lehetővé teszi nagy tömegek körében félrevezető, hamis, illetve lejáró információk terjesztését. A Hezbollah Libanonban például úgynevezett média táborokat üzemeltet, amelyekben jelentős pénzösszeg ellenében a legkorszerűbb technikák és technológiák segítségével, úgynevezett média harcosokat

képeznek, akik aztán a térségben a közösségi média mestereiként, egyenként egyszerre több száz felhasználói profilt kezelve a megrendelőik utasítására lejáratókampányokat, dezinformációs akciókat indítanak vezető helyi elöljárók, politikai érdekcsoportok, illetve politikusok lejáratására. A tehetősebb politikusoknak, illetve haduraknak saját digitális hadseregük van (Crisp & al-Salhy, 2020), (Besenyő, Gulyás & Trifunovic, 2022), (URL7).

2. Válságreagálás
A közösségi média felhasználói gyakran posztolnak katasztrófák, balesetek, terrortámadások vagy más jellegű válságok helyszínén készített képeket, információkat osztanak meg, személyes benyomásokat posztolnak. Ezeknek az információknak a meglevő ismeretekkel történő összevetését követően pontosabb következtetéseket lehet levonni a helyzetről, hol, mi történt, milyen hatás várható, hol van szükség további erőforrásokra, hogyan reagál a társadalom, vagy más országok hogyan viszonyulnak a történtekhez, milyen kommunikációt kell folytatni, kell-e változtatni a stratégiánkon?

3. A terrorizmus és szélsőségek elleni küzdelem
A terrorista szervezetek az internetet toborzásra, pénzgyűjtésre, tervezésre-szervezésre és koordinációra, valamint propaganda célokra használják. A népszerű közösségi oldalak törvényhozói nyomásra cenzúrázzák a terrorista jellegű posztokat, bejegyzéseket, azonban ennek kivédésére az elkövetők taktikát váltottak és a közvetlen propaganda helyett támogatóik megtanultak alkalmazkodni a közösségi szabályokhoz, és burkolt, közvetett propagandát folytatnak, továbbá tevékenységüket áttették az olyan végpontok közötti titkosítást lehetővé tevő platformokra, mint a Telegram, Elements, Signal stb. A szélsőséges csoportokkal hasonló a helyzet. A közösségi média cenzúráját kivédendő cenzúramentes, úgynevezett „Social Niche”-ekre költöztek. Ezek olyan közösségi platformok, mint a Gab vagy a Mastodon, amelyeken zavartalanul kommunikálhatnak, nincs cenzúra, pontosabban öncenzúrát gyakorolnak, vagyis csak a csatorna profiljába illő posztokat engedélyeznek. A dark weben ugyancsak számos szélsőséges fórum található, úgymint az Endchan, 8chan, 4chan stb.

4. Kiberbiztonság
A kiberbiztonságot napjainkban számos tényező veszélyezteti. A vírusok, kártékony programok, az egyre népszerűbbé váló zsaroló programok, a kiber kémkedés, adatlopás, az elosztásos túlterheléses támadások (DDoS) stb. Ezek elkövetői általában bűnözök, aktivisták, nemzetállamok vagy azok proxy szervezeti, illetve terrorszervezetek. A Hezbollah nevéhez például az utóbbi évtizedben

több, nagyszabású kiberkémkedési akció is köthető, amelynek során közel egy évtizeden keresztül szerte a világban távközlési, kutató-fejlesztő laboratóriumokat, felsőoktatási intézményeket, tink-tankeket támadtak meg. Az akciójuk során megszerzett információk mennyiségét még megbecsülni sem lehet (ClearSky Research Team, 2021).

Egyre inkább célponttá válik a közigazgatás, vagy a pénzügyi és ipari szervezetek. A kritikus infrastruktúra elemei különösen kedvelt célpontjai az ilyen támadásoknak. Az ellopott adatok a különböző dark webes piacokon kötnek ki, ahol bárki megvásárolhatja például izraeli haditechnikai cégektől ellopott adatokat, vagy Indonézia teljes szavazókorú lakosságának adatait, hogy csak néhányat említsek az elérhető lopott információk köréből.

A támadások megszervezése sok esetben dark webes hacker fórumokon, Telegram csatornákon történik. Ebből adódóan ezeknek a forrásoknak a figyelemmel kísérése lehetőséget biztosíthat a felkészülésre, esetleg a támadás meghiúsítására, a platformok figyelemmel kísérésével új módszereket lehet megismerni, a kiszivárgott adatok nyilvánosságra kerülése miatt keletkezett károk csökkentésére nyílhat mód.

5. Földrajzi, geopolitikai információk gyűjtése

A nyílt forrásból származó földrajzi, geopolitikai információk lehetővé teszik meglevő ismeretek pontosítását, új, a biztonsági szempontunkból fontos külföldi létesítmények felderítését, felmérését, kapacitás, illetve potenciál megbecslését, megközelítési lehetőségek felmérését stb. A közösségi médiában posztolt fényképek, videók metaadatai fontos információkkal szolgálhatnak az adott esemény helyszínével kapcsolatban. A fényképek, videók elemzése kapcsán lehetőség nyílhat személyek, hadrendi elemek, csapatmozgások, helyszínek beazonosítására.

6. Nemzetközi együttműködés

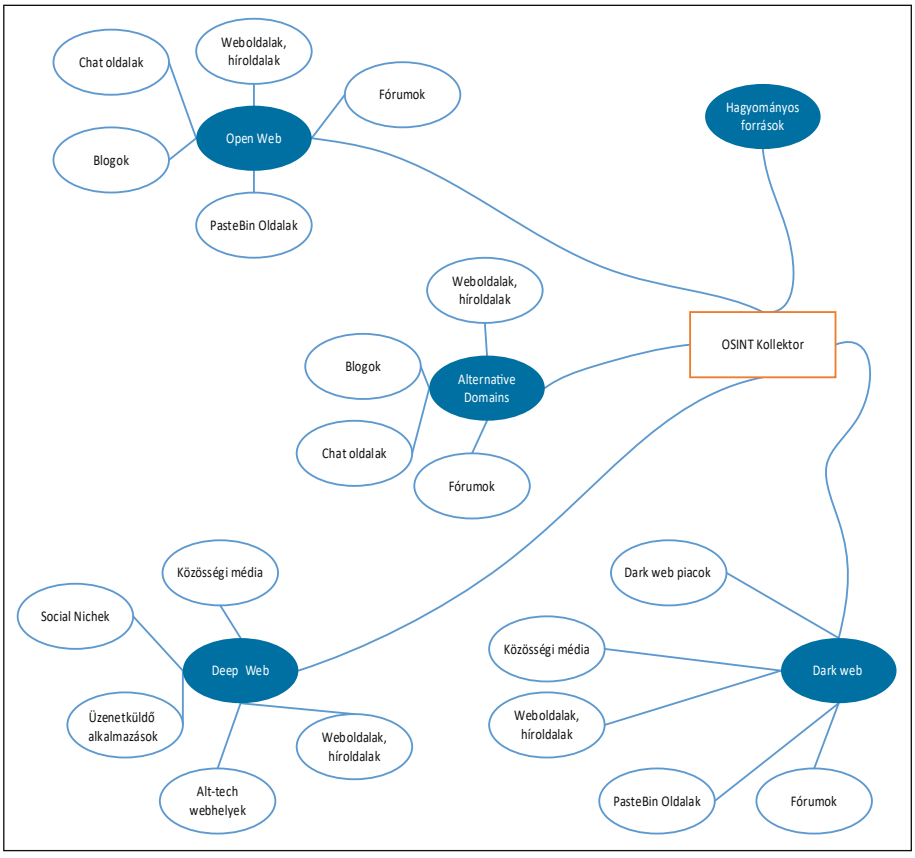
A nyílt forrásból származó információ szükség esetén megosztható a partnerszolgálatokkal anélkül, hogy a forrás biztonságát veszélyeztetné. Ezzel lehetőség nyílik a partnerszolgálatokkal való együttműködés fejlesztésére, bizonyos keretek között a kölcsönös információcsere lebonyolítására (URL5).

Az OSINT-tal elérhető források

A források tételes felsorolása kimerítené e tanulmány kereteit, ezért csak a főbb kategóriák ismertetésére szorítkozom. A lehetséges forrásokkal kapcsolatban

részletesebben foglalkozott Szabadföldi István *A mesterséges intelligenciával támogatott nyílt információszerzés (OSINT) – evolúció és kihívások* című tanulmányában (Szabadföldi, 2022), valamint Solti István *Az OSINT információgyűjtő eszközeiről* című írásában (Solti, 2019). A lehetséges internetalapú források számának és a feladat bonyolultságának érzékeltetésére készítettem az 2. számú ábrát, azonban a lista természetesen nem teljes, valamint az egyes kategóriák között átfedések is lehetnek.

2. számú ábra
OSINT források

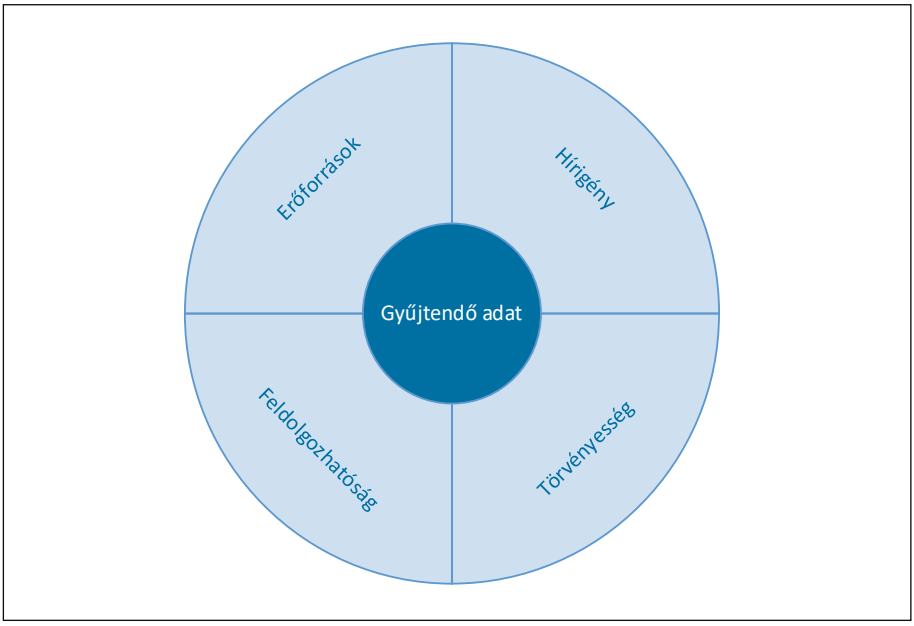


Forrás: Az ábra a szerző saját szerkesztése.

A források kiválasztásának szempontjai

Mint az már több alkalommal is említésre került, az elérhető források spektruma rendkívül széles, ez azonban nem jelenti azt, hogy mindegyikre szükség lenne. A források kiválasztása része a felderítési ciklus dinamikusan változó folyamatának. Ezek a források az aktuális szükségleteknek megfelelően változhatnak, cserélődhetnek, amihez a humán erőforrás közreműködése, az előzetes élőerős felderítés/adatgyűjtés elengedhetetlen. Mindezek mellett is a megfelelő források kiválasztásánál kompromisszumot kell kötni, hiszen az erőforrások személyi és technikai vonatkozásban is végesek. Továbbá az adatgyűjtésnek eleget kell tennie a törvényi előírásoknak, úgymint az adatgyűjtés célhoz kötöttsége (hírigény), valamint a külső engedélyhez kötött információgyűjtés engedélyezési körébe tartozó adatok gyűjtésének kizárása. A felsorolt feltételeken túl figyelembe kell venni a gyűjtendő adatok feldolgozhatóságát, amely magába foglalja azok begyűjtésének lehetőségét, tárolásukat, majd az adatok konvertálását, fordítását, homologizálását, csoportosítását, szűrését, összevetését, kereshetővé tételét. A feladat komplexitását a 3.számú ábra szemlélteti.

3. számú ábra
Kompromisszum a gyűjtendő források, illetve adatok kiválasztásánál



Forrás: Az ábra a szerző saját szerkesztése.

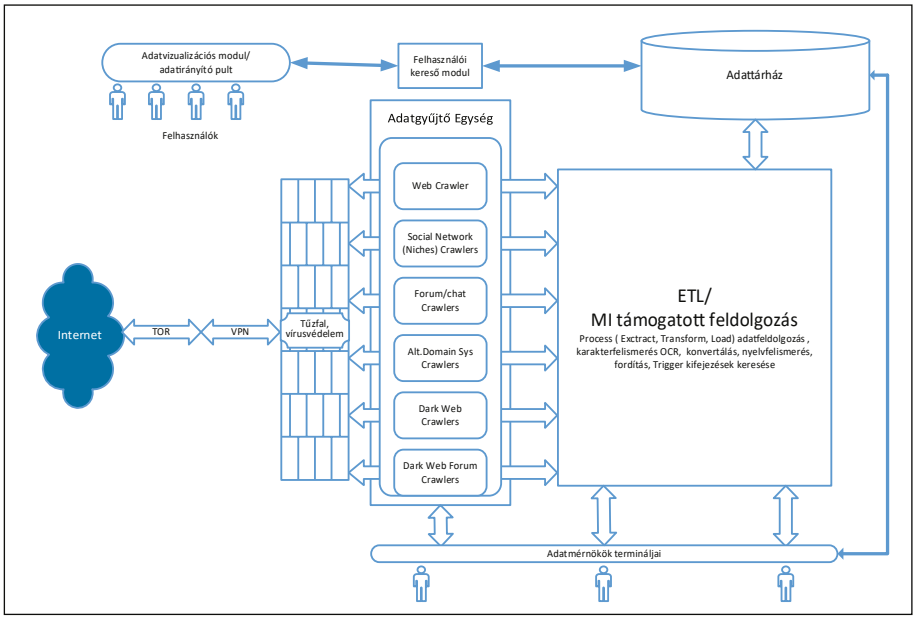
A félautomatizált nyílt forrású információgyűjtő rendszer elméleti felépítése

A félautomatizált OSINT megoldásnak az alábbi követelményeknek kell megfelelnie:

- biztosítsa az információgyűjtő szerv kilétének elfedését,
- biztosítsa az adatgyűjtést a kereskedelemben kapható eszközök által le nem fedett területekről,
- legyen felhasználóbarát,
- tegye lehetővé a felhasználói beavatkozást, az új források kijelölését, a szükségtelenek leválasztását,
- legyen moduláris felépítésű, tegye lehetővé az új források, formátumok integrálását,
- használja ki a mesterséges intelligencia és a gépi tanulás lehetőségeit,
- biztosítsa az adatok konvertálhatóságát, összevethetőségét,
- legyen képes az adatok vizualizálására,
- biztosítsa nagy mennyiségű strukturált és strukturálatlan adat tárolását, előhívhatóságát, hitelességét, selejtezhetőségét, a felhasználói tevékenység naplózását.

A fenti feltételeknek megfelelő félautomatizált adatgyűjtő rendszer elméleti felépítése az 4. számú ábrán látható.

4. számú ábra
A nyílt forrású félautomata adatgyűjtő rendszer elvi felépítése



Forrás: Az ábra a szerző saját szerkesztése.

A rendszer működése és felépítése

Rendeltetése: a felhasználók által előzetesen kiválasztott és a rendszerbe betáplált, illetve az adatgyűjtés során önállóan felderített forrásokból célirányosan fejlesztett adatgyűjtő modulok által automatikusan, inkrementált módon összegyűjtött adatok megfelelő tisztítását és konvertálását követően támogassa az összegyűjtött adatok igény szerinti információkká történő konvertálását, azok a felhasználók számára érthető és könnyen áttekinthető formában történő megjelenítését, valamint szükség szerint bizonyos információk, események online nyomon követését.

Működési elv: a rendszer a betáplált, illetve az önmaga által gyűjtött linkeket periodikusan felkeresi, az ott talált adatokat hitelesíti, növekményesen lementi, majd a begyűjtött adatokat megtisztítja, a szükséges adatfeldolgozó feladatokat – beleértve a fordítást, konvertálást, karakterfelismerést – kulcsszavak alapján történő címkézést követően adatbázisokba menti, ahonnan az adatok a napi munkavégzés céljából adattárházba kerülnek, amelyből a felhasználók jogosultságuknak megfelelően különböző szempontok alapján az adatokat előhívhatják.

Kommunikációs konfiguráció

A rendszer és az internet közötti kapcsolatnak, mint arra már korábban utaltam, biztosítania kell a gyors és anonim összeköttetést. Ennek megvalósítása természetesen több módon is lehetséges. Az általam felvázolt lehetőségek csak lehetséges megoldások, az 5. számú ábrán látható kialakítás egyfajta változat, elvi lehetőség. A képen látható megoldásnak az előnye, hogy amennyiben a rendszer egy anonim virtuális magánhálózaton keresztül csatlakozik a TOR rendszerhez, és azon keresztül kommunikál a forrásokkal a nyílt vagy mély weben és a dark weben, akkor abban az esetben az adatgyűjtő rendszerek felfedése és a felhasználói háttér beazonosításának esélye technikailag rendkívül kicsi. Ennek a megoldásnak a hátránya azonban a TOR rendszer csekély sávszélességében rejlik. A multimédiás tartalmak, különösen a nagyméretű fájlok letöltése rendkívül időigényes és biztonsági szempontból is kockázatos. Elképzelhető lehet egy másik megoldás, ahol a nyílt és a mély webet ugyan csak anonim virtuális magánhálózaton keresztül, esetleg anonim proxy szerver felhasználásával vagy annak kihagyásával szervezik, azonban a dark web irányába történő kommunikáció esetén a TOR használata ebben az esetben is megkerülhetetlen. Mindazonáltal a rendszer által felhasználástól függően generált nagyméretű adatforgalmat csökkentendő adott esetben mérlegelni kell különböző konfigurációval tervezett kettő vagy több kapcsolat párhuzamos kialakítását.

Adatgyűjtő egység

Az adatgyűjtő modulok olyan szoftverkomponensek, amelyek a különböző típusú forrásokra optimalizáltak. Ez nem zárja ki azt, hogy egy adott komponens több forrásból származó adatgyűjtésre is alkalmas legyen, azonban a tervezésénél célszerű figyelembe venni a modulrendszerből adódó előnyöket, amelyek egyszerűsítik a források változásainak követését, hiszen a biztonsági helyzet változásával új formátumú forrásokból szerzett adatokra lehet szükség a régiek megtartása vagy esetleg eldobása mellett, nem beszélve a digitális technológia gyors változásának lehetőségéről. Tapasztaljuk, hogy naponta jelennek meg új formátumú fórumok és egyéb webtartalmak.

Az adatgyűjtő egység legfőbb komponensei a crawlerek, vagy másnéven robotok, amelyek olyan szoftvermegoldások, amelyek a World Wide Weben linkeket követve weboldalakat keresnek fel, azok tartalmát lementik, az ott található releváns linkeket kinyerik, majd azokat egy meghatározott mélységig követik, aztán visszatérve az oldalra újabb linket követnek és így tovább „vándorolnak" a világhálón. Az erőforrások kímélése és a hatékonyság növelése érdekében

megadhatók olyan paraméterek, amelyekkel bizonyos oldalak vagy tartalmak, esetleg nyelvek kizárhatók a robot adatgyűjtő munkájából. A keresőóriások robotjai hasonló módon járják a World Wide Webet és indexelik a felkeresett weboldalakat (Chen, 2012).

A crawlerek a nyílt forrású információgyűjtés keretében alapvetően két célra használhatók fel, azonban természetesen lehetséges a két működési mód kombinációja is.

Felderítés: Ebben az esetben előre meghatározott keresési feltételek meghatározása mellett egy adott témában a kezdő linkek megadását követően a robotot mintegy szabadon engedve új, eddig ismeretlen weboldalak, illetve tartalmak felkutatására használjuk.

Adatgyűjtés

A robot előre meghatározott webhelyeken található tartalmak összegyűjtését, illetve a bekövetkezett változások nyomon követését végzi.

Az adatgyűjtés módszerét tekintve a robot működhet periodikus módban, amikor a megadott webhelyeket bizonyos időszakonként felkeresi és az ott található tartalmakat vagy teljes egészében, vagy pedig inkrementált módon (csak a frissítéseket) menti le. Az inkrementált adatgyűjtés kétségtelen előnye a tárolókapacitás hatékonyabb kihasználása és az adatgyűjtés idejének lerövidülése.

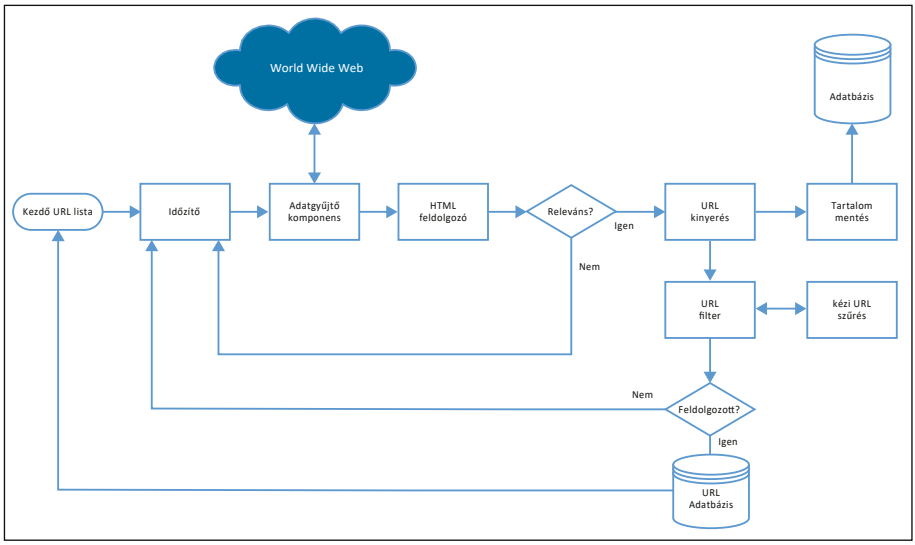
A robot másik módja a folyamatos működés, amikor a megfigyelt webhellyel folyamatos kapcsolatban van és a változásokat inkrementálva menti.

A különböző működési módok alkalmazása függ a szükséges információk jellegétől, illetve a megfigyelt forrás sajátosságaitól.

Ebben az összefüggésben a robotok lehetnek weboldal, fórum, közösségi média és chat adatgyűjtő robotok. Ezek kialakítása és működési sajátosságai függenek az információforrás elérhetőségétől, struktúrájától, illetve a célhelyen tárolt információ formátumától. A 5. számú ábrán egy weboldalakat felkereső robot elvi felépítése látható.

5. számú ábra

WebCrawler (webrobot) működése

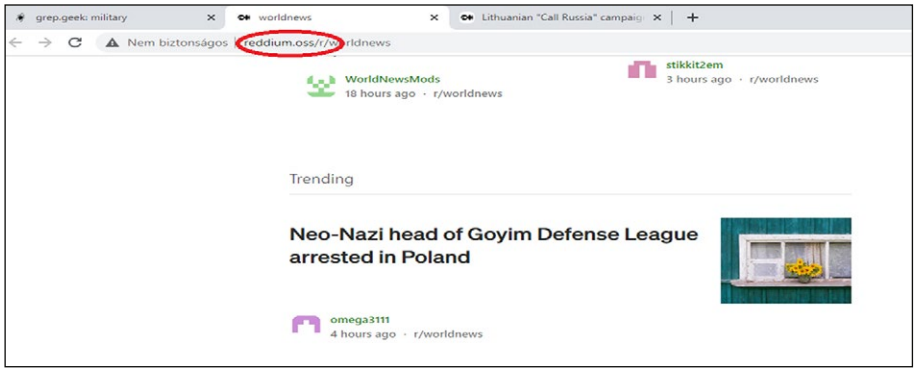


Forrás: Az ábra a szerző saját szerkesztése.

A különböző alternatív domaineken, illetve a dark weben működő robotok kialakítása lényegében megegyezik a hagyományos yyílt interneten használt szoftvermegoldásokkal, mindössze a célhelyek elérésének módja különbözik. A 6. számú ábrán egy alternatív domainen található weboldal képernyőfotója látható.

6. számú ábra

Példa alternatív domainen található weboldalra



Forrás: Az ábra a szerző képernyőfotója.

Adatgyűjtés a dark weben

Mint arra már korábban utaltam, a dark weben történő adatgyűjtésre történő előzetes felkészülés összehasonlítva a nyílt interneten történő felkészüléssel több időt, körültekintést és előkészítést igényel.

A sikeres adatgyűjtés feltétele a megfelelő források kiválasztása, fokozottan igaz ez a dark web területén, ahol gyakran keresési találatokat követve félrevezető linkek terelik el a kutatót olyan tiltott tartalmak irányába, mint például a gyermekpornográfia, ezért az előzetes felkészülés során figyelembe kell venni a dark web sajátosságait, amelyek a következők.

1) A dark webes környezeten belül a TOR és az I2P rendszerben a tartalom létrehozójának és felhasználónak a személyazonossága és földrajzi tartózkodási helye rejtve marad.
2) A dark web működésének sajátosságaiból adódóan a rendszeren nem működik a nyílt internetes keresők képességével összemérhető kereső rendszer. A weboldalak üzemeltetői oldalaikat különböző linkgyűjteményeken hirdetik (HiddenWiki, TOR links stb.), illetve a rendszeren működő keresőmotoroknál regisztrálják, amennyiben az oldal elérhetőségét nyilvánosságra kívánják hozni. Ellenkező esetben az oldal készítőjén kívül másnak nincs tudomása az oldal létezéséről.
3) A TOR és az I2P rendszereken bármely felhasználó, akár a saját számítástechnikai eszközén is, tetszőlegesen létrehozhat webszolgáltatást, amely az eszköz kikapcsolásával elérhetetlenné válik. Ennek következtében nagy az oldalak fluktuációja, és elérhetőségük gyakran teljesen kiszámíthatatlan. A Freeneten létrehozott tartalmak nem függnek a létrehozó számítástechnikai eszközeinek állapotától, ezzel szemben a ritkán látogatott oldalak – az erőforrások kímélése céljából – eltűnnek a rendszerből, és csak a létrehozó általi újbóli aktiválását követően válnak ismét elérhetővé. Az eltűnés várható ideje megjósolhatatlan, hiszen szorosan függ a rendszert használók létszámától és az általuk generált forgalomtól. Ezért a Freenetes tartalmak elérhetősége szintén kiszámíthatatlan.
4) A dark webes kommunikáció többszörös titkosítási eljárással védett, ez azonban azzal jár, hogy a kommunikációs sávszélesség lecsökken, a puszta navigálás, vagy a nagyméretű fájlok letöltése is időigényes.
5) A weboldalak külső megjelenése és funkciógazdagsága a TOR és az I2P rendszeren szegényesebb annak köszönhetően, hogy biztonsági okokból a JavaScript alapú programok tiltásra kerülnek, ugyanis az ilyen scriptek segítségével, a „browser fingerprinting” technológiát alkalmazva, a felhasználók beazonosíthatóságának kockázata megnőne.

Az előzetes adatgyűjtés érdekében igénybe vehető források

Az adatgyűjtést célszerű a nyílt interneten a már korábban jelzett anonimizáló megkötésekkel kezdeni. Számos olyan linkgyűjtemény érhető el, amelyek tematikus csoportosításban tartalmazzák a legnépszerűbb, illetve legismertebb dark webes szolgáltatások linkjeit, amelyeket aztán a dark weben követve további gyűjteményeket érhetünk el.

Információforrások felkutatása céljából igénybe vehetők olyan dark webes keresők, mint például a Phobos, Haystack, Ahmia, azonban ezeken jóval kevesebb releváns találat érhető, mint azt a nyílt webes keresőknél megszoktuk.

A dark webes tartalmak sajátosságai

Az automatizált adatgyűjtés szempontjából a dark weben található szolgáltatások formátumai főbb vonalaikban megegyeznek a nyílt webes szolgáltatások jellemzőivel, ezért a dark webre tervezett robotoknak sem kell jelentős mértékben eltérniük. A lényeges különbség a csatlakozás módja, amely speciális konfigurációt igényel a TOR, az I2P, illetve a Freenet esetében.

A TOR rendszer 2021 júliusától a nagyobb biztonság érdekében áttért a 16 karakteres V2. verziójú címzési rendszerről az új 56 karakteres V3. verzióra. A gyakorlatban a felhasználó számára ez azt jelenti, hogy a korábbi 16 karakteres címeken található tartalmak a legfrissebb szoftverrel már nem érhetők el, azonban ezek az oldalak még létezhetnek és felhasználhatók lehetnek mint információforrások. Az elérésükhöz azonban szükség van a TOR rendszer régebbi verziójára, amely még kezelte ezt a címzési módot. Az előnyöket és a hátrányokat mérlegelve kell eldönteni, hogy az így elérhető adatforrások bevonhatók-e az adatgyűjtésbe.

Mesterséges intelligencia és adatfeldolgozás

Az adatgyűjtő modulok által begyűjtött adatok feldolgozása nem nélkülözheti a korszerű technológiák és módszerek alkalmazását, ugyanakkor az emberi közreműködés még a legkorszerűbb technológiák alkalmazása esetén is kikerülhetetlen.

A mesterséges intelligencia (MI) felhasználhatóságának lehetőségei

A körülöttünk zajló világ egy új kulcsszótól, a mesterséges intelligenciától hangos. Napjaink varázs szava a mesterséges intelligencia. Vizsgáljuk meg mindenek előtt, hogy mi is ez önmagában?

A mesterséges intelligencia, vagy angolul Artificial Intelligence (AI) tulajdonképpen egy számítógépes rendszer azon képessége, hogy emberihez hasonló kognitív folyamatokat tud utánozni, amilyen például a tanulás vagy problémamegoldás (URL1).

A már említett számítógépes rendszer a matematika és a logika felhasználásával képes utánozni azt az érvelési eljárást, amit az emberi elme végez, amikor új információk alapján tanul vagy döntést hoz (URL1).

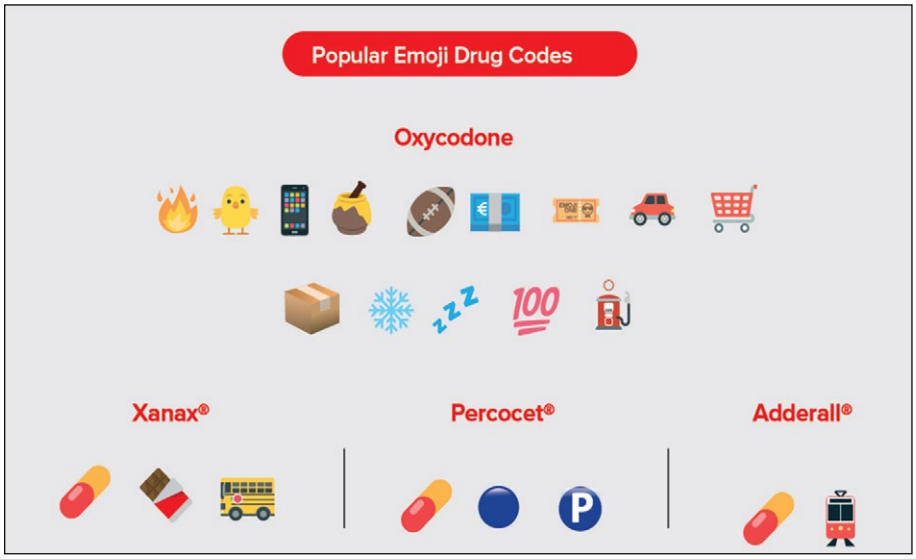
Az MI-t használó rendszerek képesek előrejelzéseket adni, a meglevő adataiból származtatott mintázatokon alapuló feladatokat hajtanak végre, továbbá képesek tanulni a saját hibáikból, így növelve pontosságukat. Egy kellőképpen betanított (kifejlett) MI az új információkat rendkívül gyorsan és pontosan dolgozza fel, lásd például képfelismerő programokban történő felhasználásukat (URL1).

Az MI korlátai

A jelenlegi MI rendszerek – mint azt említettük – a szűk mesterséges intelligencia (Narrow Artificial Intelligence, NAI) körébe tartoznak, amelyek nem képesek emberi módon gondolkodni, csak egy speciális feladat végrehajtására optimalizáltak, azt azonban az embernél sokkal gyorsabban képesek végrehajtani. Ezek a rendszerek könnyen megtéveszthetők, kijátszhatók. A cenzúrát végző MI megtévesztésére elterjedt egyik módszer az üzenetekben a tiltott szavak emojikra (a szöveges üzenetbe illeszthető rajzok) cserélése, amelynek segítségével a drogkereskedők, vagy a tiltott tartalmat megjelenítők ki tudták játszani az automata ellenőrző rendszert, és zavartalanul folytatják illegális tevékenységüket.

Az amerikai Drug Enforcement Administration (DEA) útmutatót is adott ki a közösségi médiában folytatott drogterjesztés során alkalmazott emoji kódokról, amelyek segítségével a gépi algoritmusokat megtévesztették (Drug Enforcement Administration, 2021), (URL8). A kiadványuk a 7. számú ábrán látható.

7. számú ábra
Kábítószer megnevezést helyettesítő emojik a közösségi médiában



Forrás: OPCK_2.0_Emoji Codes-Poster.

Elterjedt módszer a képek kismértékű elváltoztatása, amely lehetetlenné teszi a képek útjának technikai nyomon követését, vagy egyes tiltott szavak helyett virágnyelvű kifejezések használatát. Például az Iszlám Állam hivatalos folyóiratára, az Al-Nabara „A Muszlim Újság”-ként utalnak (Meili, 2022). Az MI-t az ilyen és ehhez hasonló trükkökre felkészíteni nem lehet, ezért könnyen belátható, hogy ezek a rendszerek viszonylagos önállóságuk mellett is igénylik a szoros emberi felügyeletet.

Az MI integrálása az adatgyűjtő rendszerbe

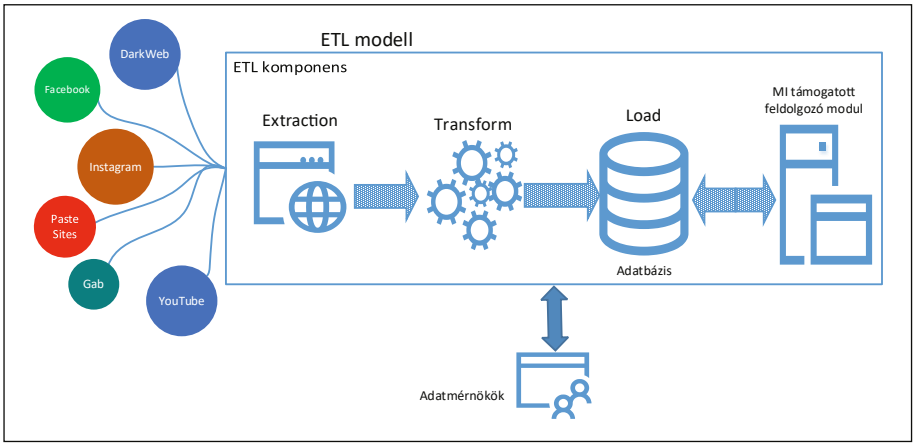
Az MI alkalmazásának előkészítése során a modellt gépi tanulás szakértőnek kell felépítenie, majd az összegyűjtött adatok tisztítását, rendszerezését adatmérnökök végzik, azonban a megfelelő adatok kiválasztásához, a leggyakoribb lekérdezések lehetséges változatainak kidolgozásához bűnügyi, illetve nemzetbiztonsági szakemberek közreműködésére van szükség. Végül az elkészült, működő rendszert a megfelelő biztonsági és titokvédelmi rendszabályok betartása mellet integrálni kell a szervezet meglévő folyamataiba, informatikai környezetébe, ezért alkalmazásfejlesztők bevonása is szükséges.

Adatfeldolgozás

ETL modell (Extract Transform Load)

Az automatizált adatgyűjtés során az összegyűjtött adatok nyelvüket, formátumukat tekintve rendkívül változatosak lehetnek, ezért ezek feldolgozhatóvá tétele, egységesítése, tisztítása, a duplikátumok eltávolítása rendkívüli jelentőséggel bír a későbbi feldolgozó, elemző munka eredményes végrehajtása érdekében. Az összegyűjtött információk szövegbányászati eszközökkel való feldolgozásra alkalmassá tételére az úgy nevezett ETL modul szolgál. Itt történik a szükséges adatok kivonása, átalakítása és átmeneti adatbázisba töltése. A 8. számú ábrán egy ilyen modulnak az elvi felépítése látható.

8. számú ábra
Az ETL rendszer elvi felépítése



Forrás: Az ábra a szerző saját szerkesztése.

Az adatátalakítás olyan műveleteket foglal magába, mint a különböző szöveges dokumentumok egységes szövegformátumba konvertálása, a nyelv felismerése és fordítása. Az optikai karakterfelismerés segítségével a különböző képi formátumokban, illetve a „Speech to Text" technológiával a videókban és a hangállományokban található szöveges állomány kinyerése és szöveggé alakítása. Az átalakítás során az ETL modul igénybe veszi az MI támogatott feldolgozó modul képességeit, amelyekre a későbbiekben részleteiben kitérünk.

A feldolgozott állományok átmeneti tárolóba kerülnek, ahonnan a feldolgozó modulba folytatják útjukat.

A három folyamat természetesen futhat párhuzamosan, hiszen az adatok kivonásával egy időben a már korábban kivont adatok feldolgozása akadály nélkül folyhat, miközben a már feldolgozott adatokat tárolóba helyezheti a rendszer. Ezzel a módszerrel jelentős időmegtakarítást lehet elérni (Raunak, Jhawar & Tejada, 2022), (URL9).

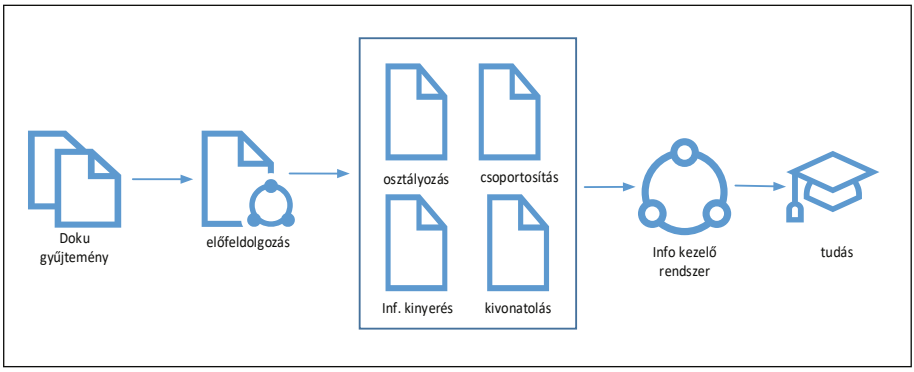
Feldolgozó modul

Szövegbányászat

A szövegbányászat a számítástudomány azon része, amely az elektronikus dokumentumok feldolgozását és elemzését végzi. Napjainkban ilyen dokumentumok keletkeznek az élet minden területén, és olyan mennyiségben fordulnak elő, hogy feldolgozásuk meghaladja az emberi teljesítőképesség határait. Az ügyek intézése, az információk továbbítása nagyrészt elektronikus dokumentumokon keresztül történik, amelyek lehetnek emailek, word dokumentumok, emlékeztetők, PDF dokumentumok, weblapok, közösségi média üzenetek stb., tehát összességében strukturálatlan adatok. Az MI-veltámogatott szövegbányászat ennek a nagy mennyiségű szöveges tartalomnak a kezelésében, a közöttük való keresésben, illetve a bennük rejlő, addig nem ismert értékes, rejtett összefüggések feltárásban nyújt segítséget (Tikk, 2007).

Az emberi nyelv nem a számítógépes feldolgozás szempontjai, hanem a szóbeli vagy írásbeli kommunikáció érdekében alakult ki. Az emberek könnyen azonosítják a stílust, a kontexust, és alkalmazzák a nyelvi mintákat, ezek azonban a számítógépek számára egyes esetekben szinte megoldhatatlan feladatnak számítanak. Az emberi elme képes felfogni és megérteni a strukturálatlan szövegeket, azonban nem vagyunk képesek nagy mennyiségű szöveg gyors feldolgozására. A szövegbányászat lényege az ember nyelvi tudásának ötvözése a számítógépek gyorsaságával. A szövegbányászati feladatok megoldásakor több tudományterület vívmányait is igénybe kell venni. Szükség van matematikai, informatikai, gépi tanulással kapcsolatos, valamint a számítógépes nyelvészet és a természetes nyelvek feldolgozásával kapcsolatos eszközök használatára (Tikk, 2007). A 9. számú ábrán egy szövegbányász rendszer elméleti felépítése látható.

9. számú ábra
Szövegbányász rendszer elméleti felépítése



Forrás: Tikk, 2007.

A folyamat során az ETL modul által előállított egységesített szöveges tartalom kerül feldolgozásra, amelynek keretében a szövegeket adatokká konvertálják. A cikknek nem feladata a szövegbányászati eljárások bemutatása, hiszen ezek ismertetése meghaladná e tanulmány kereteit, így most csak arra szorítkozom, hogy a technológia segítségével kinyerhető adatok jellegét felsoroljam.

A kellően előkészített és különböző szövegbányászati eljárásokkal feldolgozott szövegből neveket, helyszíneket, kapcsolatokat lehet azonosítani. Szövegbányászati módszerekkel lehetőség van a szerző számos személyes tulajdonságának megismerésére, úgymint iskolázottság, nyelvismeret, műveltség, adott esetben foglalkozás vagy hobbi valószínűsítésére. Ismeretlen szerzős szöveg esetén a meglévő szöveges adatbázist felhasználva a szerző kilétének megállapítására. Ezzel a módszerrel az is felderíthető, ha a hozzászóló különböző profilokat felhasználva tevékenykedik egy vagy akár több webhelyen is váltogatva az identitását.

Természetes Nyelvi Feldolgozó modul (Natural Language Processing, NLP)

Természetes nyelveknek az emberek, közösségek által használt olyan nyelvet nevezzük, amely az idők során vagy tudatos nyelvújítás útján, vagy spontán folyamatosan változik, fejlődik.

A természetes nyelvi feldolgozás (a továbbiakban NLP) az MI felhasználásával képessé teszi a számítógépes rendszereket az írott vagy a beszélt szöveg emberhez hasonló módon történő megértésére, illetve feldolgozására (URL2).

Az NLP több tudományág együttműködéséből jött létre, hiszen magába foglalja a számítógépes nyelvészetet (vagyis az emberi nyelv szabály alapú modellezését),

továbbá különböző gépi tanulási módszereket, beleértve a mély tanulást és a statisztika tudományát. A technológiák együttműködéseként lehetővé válik a számítógép számára hang vagy szöveges formában az emberi nyelv feldolgozása, valamint annak megértése, kiegészítve a beszélő szándékával, illetve érzelmeivel (URL2).

Ez a technológia képes szövegek automatikus fordítására, megérti a szóbeli parancsokat (zavaró környezetben is), alkalmas nagy mennyiségű szövegek valós idejű összegzésére, kivonatolására, illetve szöveg beszéddé vagy beszéd szöveggé történő átalakítására.

A következőkben nézzük milyen feladatokra alkalmas az NLP technológia.

- Beszédfelismerés, parancsértelmezés vagy Speech-to-Text feladata: az ilyen rendszerek, mint például az Otter (https://otter.ai/) webszolgáltatás rendkívüli pontossággal képesek a zavaró háttérzajok, akcentusok, nyelvtani hibák, beszédzavarok mellett is a beszédet szöveggé átalakítani.
- Nyelvtani szabályok alkalmazása (grammatical tagging): képesek a különböző szófajokat, mondatrészeket azonosítani.
- Több jelentésű szavak környezettől függő helyes értelmezésére.
- Szavak értelmük szerinti csoportosítása. Különbséget tudnak tenni például földrajzi nevek és személynevek között.
- Szentiment elemzés, amelynek keretében a szövegeket tartalmuk alapján pozitív, negatív vagy semleges kategóriába sorolják a szerzőjük mondanivalójától, illetve hangulatától függően.
- Természetes nyelvi szöveg előállítása a beszédfelismerés ellenpárja (URL2).

Mindezen képességek az alábbi területeken használhatók fel a gyakorlatban.

- A polgári felhasználástól (Spam Detection) eltérően az adatgyűjtő rendszerben ez a képesség megfelelő programozást követően hatékonyan használható fel a beérkező információk téma szerinti csoportosítására, a kulcs kifejezések, trigger események kiválasztásra.
- Gépi fordítás (Machine Translation). A mindenki által ismert Google fordító az egyik népszerű képviselője ennek a technológiának. A gép nem egyszerűen tükörfordítást hajt végre, hanem felismeri az egyes kifejezéseket, szólásokat is. A gépi fordítás elengedhetetlen eleme az adatgyűjtő rendszernek.
- Közösségi média hangulatelemzés (Social Media Sentiment Analysis), illetve érzelmi elemzés. Az NLP segítségével lehetőség van a szerzői hozzáállás nyelvi elemeinek keresésével automatizáltan megállapítani, hogy az adott szöveg, poszt szerzője, vagy a beszélő milyen hangulatban volt, mi a véleménye a világról. A szerzőtől származó több szöveg esetén, időrendet

felállítva meg lehet állapítani az esetleges radikalizálódását, jóslatot lehet tenni arra, hogy várható-e részéről erőszakos megnyilvánulás. Az ilyen következtetések a szóhasználatából, az általa használt nyelvtani szerkezetekből vonhatók le. Ezek a jellemzők ugyancsak segíthetnek az ismeretlen szerzők beazonosításában is (Drávucz, Szabó & Vincze, 2017).

- Kivonatolás, összegzés. Az összegyűjtött adatokból különböző szempontok alapján lehetséges kivonatok, összegzések készítése, amelyek hozzájárulhatnak egy adott témakör jobb megértéséhez, illetve bemutatásához.

Képfelismerés (Computer Vision, CV)

A Computer Vision az a módszer, ahogy az MI megtanulja látni a körülötte lévő világot. A képfelismerés vagy objektumfelismerés olyan technológia, amely a gépi látás segítségével lehetővé teszi az MI számára a körülöttünk lévő világ tárgyainak az emberhez hasonló módon történő felismerését és kategorizálását. A rendszer mély neurális hálók segítségével képes felismerni számos referenciakép alapján a tanítás során megtanult mintázatokat, majd ezek segítségével a képeket kategóriákba sorolni.

Ennek a technológiának a segítségével lehetőség van az összegyűjtött képek és videótartalmak elemzésére, a rajtuk, illetve bennük szereplő adatok (rendszámok, fegyverek stb.), személyek kigyűjtésére, és megfelelő adatbázis esetén még a beazonosításukra is. Geolokációs szoftverek segítségével a fényképek készítésének helye is beazonosítható lehet. A közösségimédia-elemzéssel kombinálva személyek kapcsolatai, illetve a kérdéses személlyel kapcsolatba hozható helyszínek is beazonosíthatóvá válhatnak.

Adatbázis

Adatbázison a hétköznapi értelemben valamilyen rendezett, rendszer szerint tárolt adatok rendszerét értjük, amelyet tárolásra, szerkesztésre és lekérdezésre alkalmas szoftver kezel. Az adatok pusztán nagy száma nem tekinthető adatbázisnak, csak amennyiben valamiféle rend szerint épül fel és lehetővé teszi az adatok értelmes kezelését.

Az adatgyűjtő munka során keletkezett adatokat adatbázisban kell tárolni, úgy hogy az adat sértetlensége (integritása), kikereshetősége biztosított legyen. Az interneten az adatok tőlünk függetlenül megváltozhatnak, eltűnhetnek, illetve megsemmisülhetnek, ezért a begyűjtött adatot eredeti formájában, hitelesítve is tárolni kell. A gyűjtött adatok egy része a későbbiekben akár bizonyítékként is felhasználható lehet.

Természetesen a feldolgozott adatokat is tárolni kell annak érdekében, hogy a későbbiekben is feldolgozhatók, más adatokkal összevethetők, illetve visszakereshetők legyenek.

Az adatgyűjtő munka során keletkezett adatok alapvetően három kategóriába sorolhatók.

Strukturált adatoknak nevezzük azokat az adatokat, amelyeknek elemei között valamilyen meghatározott kapcsolat áll fenn. Ezek jellemzően lehetnek excel táblák vagy egyéb táblázatok, statisztikai adatbázisok stb., amelyek elemzése viszonylag egyszerű, köszönhetően a bennük tárolt adatok strukturáltságának.

A félig strukturált adatok, amelyek valamilyen címkével meg vannak jelölve (meta tags), azonban bennük az adatok teljesen strukturálatlanok.

A strukturálatlan adatok változatos formáját tartalmazzák az adatoknak, beleértve a számokat, statisztikákat, képeket, közösségi média posztokat, hanganyagokat, videókat, véleményeket, hozzászólásokat tartalmazhatnak, amelyeknek tömeges elemzése komoly kihívást jelentene az emberi feldolgozás során. A napjainkban keletkező digitális dokumentumok többsége ez utóbbi kategóriába tartozik és ezek feldolgozásának elengedhetetlen feltétele az MI alkalmazása.

A tárolt adatok formátumuktól függően különböző típusú adatbázisokba mentendők, ugyanis jellegüknél fogva más típusú adatbázis szolgál a strukturált adatok és megint más a strukturálatlan adatok tárolására. Ennek alapján alapvetően két fajta adatbázistípusról beszélhetünk, úgymint a relációs típusú adatbázisokról (relational database management system RDBMS) a strukturált adatok tárolására, illetve a nem relációs adatbázis típusú, úgynevezett NoSQL adatbázisokról.

A strukturálatlan adatokat célszerű olyan adatbázisokban tárolni, amelyek képesek a különböző formátumú adatokat tárolni, azokat szerkeszteni, és biztosítják az adatokból történő különböző szempontok alapján történő lekérdezés lehetőségét. Erre a célra a legalkalmasabbak az úgynevezett nem hagyományos, vagyis nem relációs adatbázisok.

A NoSQL adatbázisok

Az adatbázisok elnevezése arra utal, hogy a sorokat és oszlopokat kezelő SQL lekérdező nyelvet (Structured Query Language, SQL) relációs használó adatbázisoktól eltérően gyorsan változó, nagy mennyiségű adatok kezelésére képesek. A nem relációs adatbázisok olyan tárolási modellt alkalmaznak, amely az adatoktól függő speciális követelményekre vannak optimalizálva. Az adatok tárolhatók kulcsérték párokként, dokumentumokként vagy gráfokként. Ezeknek az adatbázisoknak az erőssége az adatok tárolása és lekérdezhetősége, szemben a relációs adatbázisokkal, ahol jelentős szerepet kap a tranzakciók (az adatbázisban végzett műveletek) naplózása.

Maga a technológia nem új, hiszen már az 1960-as években is létezett, azonban igazi létjogosultságát napjainkban, a Big Data korában érte el (URL3).

Adattárház

Az adattárház (data warehouse) az információtechnológia viszonylag új területe, hiszen alig több mint egy évtizedes múltra tekint vissza. Létrejöttét hasonlóan a felhőalapú adattároláshoz és műveletekhez a korábban elképzelhetetlen mennyiségű digitális információ, illetve a műszaki fejlődés tette lehetővé.

Az adattárház egy szervezet azon adatgyűjtő és szolgáltató részeit tartalmazza, ahol a keletkezett működési (esetünkben az összegyűjtött és feldolgozott) adatokat hatékony és egyszerűen kezelhető elemzésekhez újrastrukturálják.

A különböző típusú adatbázisok kezelése és folyamatos működésének biztosítása érdekében célszerű az adatbázisokat egy úgynevezett adattárházba szervezni. Az adattárház létrehozását az indokolja, hogy az adatgyűjtő rendszerek az adatok tárolására, illetve típusuktól függően az adatbázis műveletek nyilvántartására vannak optimalizálva, ezért nem hatékonyak a nagyszámú, különböző adatból összevont lekérdezések végrehajtásában. Ezek olyan terhelést jelentenének a rendszernek, amely nem biztosítaná a feldolgozó folyamatok folyamatos kiszolgálását. Az adattárházba történő szervezés lehetővé teszi ezen rendszerek tehermentesítését, mivel ezek lekérdezésekre vannak optimalizálva. Meghatározott időközönként átveszik az adatokat a forrásadatbázisokból, és az átvett adatokból tudják kiszolgálni a kéréseket (Sidló, 2004), (URL10).

Az adattárház működése (data warehousing) három fő folyamatot foglal magába:

1) Adatkinyerés a forrásadatbázisokból.
2) A kinyert adatok formálása különböző riportok, illetve jelentések számára.
3) A riportok, illetve jelentések elérhetővé tétele a felhasználók, elemzők számára

Lekérdező modul

A lekérdező modul teszi lehetővé, hogy az adattárházból a felhasználók jogosultságaiknak megfelelően adatokat kérjenek le. A munka meggyorsítása érdekében biztosítani kell a felhasználók igényei szerint megformált és eltárolt (gyorselérésű) lekérdezések, illetve az egyedi kérések teljesítését lehetővé tevő egyedi lekérdezések futtatásának lehetőségét is. Az adatbázis jogtalan megváltoztatása elkerülésének érdekében az adatmódosító lekérdezéseket ugyancsak szigorú jogosultsági hierarchiához kell kötni. A kereső modulon végzett

felhasználói tevékenységet a felhasználók tevékenységének ellenőrzése érdekében célszerű naplózni.

Adatvizualizáció

Az adatvizualizáció lényege a felhasználó számára az adatok rendezett, átlátható formában történő bemutatása. Az adatvizualizáció tulajdonképpen életre kelti az adatokat elérhetővé téve a felhasználó számára, hogy minél rövidebb idő alatt, minél pontosabb és átláthatóbb képet kapjon a megfigyelt adatokról. Ezzel a technológiával az összes széttagolt, aprólékos adat könnyen áttekinthető, látványos, informatív formában jeleníthető meg. Az adatvizualizációs modul a felhasználó szempontjából tulajdonképpen az egyik legfontosabb modul.

Az adatvizualizációs modult a felhasználó az adatirányító pulton keresztül éri el.

Adatirányító pult

Az adatirányítópult olyan eszköz, amelyet az adatok nyomon követésére, elemzésére és megjelenítésére használnak abból a célból, hogy betekintést nyerjenek a megfigyelt folyamatok alakulásába úgy, hogy a megjelenített adatok a felhasználó számára könnyen áttekinthetők és informatívak legyenek.

A fizikai megjelenítést a felhasználók igényei alapján kell megtervezni, lehetővé téve az igényeknek megfelelő testreszabhatóságát. A lekérdezéseire adott válaszok és a megfigyelt folyamatok, események ebben a modulban válnak láthatóvá. A felhasználók igényeinek és a célszerűségnek megfelelően megtervezett megjelenítő modul a kapott adatokat szemléletessé, áttekinthetővé teszi, a meghatározott események bekövetkeztekor a felhasználók felé riasztást adhat. Ezeknél a tulajdonságainál fogva lehetővé teszi a vezetők időbeni, szemléletes tájékoztatását, elősegítve a tervezést, illetve a döntéselőkészítést. A 10. számú ábrán egy képernyőfotó látható a kanadai Flashpoint OSINT üzleti információgyűjtő vállalkozás ransomware-ek terjedését nyomon követő adatirányító pultjáról.

10. számú ábra
A Flashpoint kanadai polgári OSINT vállalkozás zsarolóprogramokkal (ransomware) kapcsolatos adatirányító pultjának képernyőfotója

Forrás: (URL11).

A hatékony adatirányító pulttal szembeni követelmények

Legyen képes együttműködni a rendszer többi elemével, az adatvédelmi követelményeknek eleget téve illeszkedjen a szervezeti szoftverkörnyezet megfelelő elemeihez.

Valós adatokat használjon: biztosítsa a felhasználó részére a hiteles, időszerű adatokkal végezhető munkát.

Felhasználóbarát legyen: biztosítsa a funkciók, színek, szűrők egységességét, hogy a felhasználó könnyen eligazodjon és áttekinthesse a rendszert.

Igazodjon a felhasználók igényeihez, biztosítsa a felhasználók különböző szerepéből adódó eltérő igényeknek megfelelő testreszabhatóságot.

Adatbiztonság és adatvédelem: jogosultsági rendszer kialakításával akadályozza meg az illetéktelen hozzáférést (URL4).

Láthatóan az adatpulttal szembeni követelmények szerteágazóak, azonban, ha ezee követelmények teljesítését követően az elemzők munkája gyorsabbá és hatékonyabbá válhat, így a vezetői döntéselőkészítéshez rövidebb idő alatt pontosabb és látványában meggyőzőbb információkkal járulhatnak hozzá.

Következtetés

Ahogyan a bevezetőben utaltam rá, napjainkban forradalom zajlik a digitális adatok, illetve információk feldolgozásának gépesítésében, illetve automatizálásában.

Ez a forradalom el fogja sodorni azokat a szervezeteket, amelyek nem képesek lépést tartani az új technológiákkal, vagy vonakodnak azokat bevezetni.

Napjainkban elmosódik a határ a fizikai világ és a digitális valóság között, ami azzal jár, hogy a biztonsági és rendvédelmi szerveknek meg kell érteniük, hogy mi történik a „terepen", és ehhez nem elég a hagyományos felderítés, hanem be kell járniuk a virtuális világ rejtett zugait is, beleértve a nyílt, a mély és a sötét webet is. A naponta keletkező óriási mennyiségű adatban nemzetbiztonsági szempontból fontos információk rejtőznek, illetve rejtőzhetnek, amelyek megtalálására a hagyományos módszerekkel egyre kisebb az esély. A szolgálatoknak, az új technológiák intuitív alkalmazásával, ezeket a rejtett információkat fel kell deríteniük, hogy eleget tehessenek a törvényben előírt kötelezettségeiknek.

A tanulmány célja egy nyílt forrású, félautomatizált adatgyűjtő rendszer kialakítási lehetőségének vizsgálata volt. A célok között szerepelt annak megállapítása, hogy egy hasonlóan kialakított rendszer mennyire képes autonóm módon, gyakorlatilag minimális emberi beavatkozással működni. Megállapítható, hogy a jelenlegi technikai fejlettségi szinten, leginkább az MI korlátai miatt az emberi közreműködés nem mellőzhető az adatgyűjtés folyamán. Az adatfeldolgozás ma még elképzelhetetlen az adatfeldolgozásban jártas adatmérnökök nélkül. Ugyanakkor kétségtelen, hogy még egy félautomatizált rendszer is nagyságrendekkel megnövelheti a begyűjthető adatok és azok feldolgozásának nagyságát, illetve annak sebességét.

A bemutatott modell csak egy elméleti megvalósítási lehetőség, ábrázolásával pusztán a megoldásban rejlő lehetőségek széles skáláját kívántam érzékeltetni, azzal a nem titkolt céllal, hogy a későbbiekben egy megvalósuló projekt alapját képezheti. A gyakorlati életben természetesen nem feltétlenül kell minden elemet egyszerre megvalósítani vagy létrehozni, hiszen már egyes részegységek működtetése is jelentősen felgyorsíthatja és hatékonyabbá teheti az adatgyűjtés folyamatát.

Befejezésül Amy B. Zegartot az elismert amerikai titkosszolgálati szakértőt szeretném idézni, aki a *Spies, Lies, and Algorithms: The History and Future of American Intelligence* című könyvében azt írja, hogy az amerikai titkosszolgálatoknak alkalmazkodniuk kell, különben elbuknak (Zegart, 2022). Azt hiszem ez a mondat ránk is érvényes.

Felhasznált irodalom

Army Techniques Publication (2012). *Open-Source Intelligence,* ATP 2-22.9. Army Techniques Publication.

Besenyő, J., Gulyás, A. & Trifunovic, D. (2022). Hezbollah and the Internet in the Twenty-First Century. *International Journal of Intelligence and CounterIntelligence*, *36*(3), 669–685. https://doi.org/10.1080/08850607.2022.2111999

Chen, H. (2012). *Dark Web: exploring and data mining the dark side of the web*. Springer. https://doi.org/10.1007/978-1-4614-1557-2

ClearSky Cyber Security Ltd. (2021). *"Lebanese Cedar" APT Global Lebanese Espionage Campaign Leveraging Web Servers*. ClearSky Cyber Security Ltd.

Drávucz F., Szabó M. K. & Vincze V. (2017). Szentiment- és emóciószótárak eredményességének mérése emóció- és szentimentkorp. In Vincze V. (Szerk.), *XIII. Magyar Számítógépes Nyelvészeti Konferencia* (pp. 228–239). Szegedi Tudományegyetem Informatikai Tanszékcsoport.

Hobbs, C., Moran, M. & Salisbury, D. (2014). *Open source intelligence in the twenty-first century: new approaches and opportunities*. Palgrave Macmillan. https://doi.org/10.1057/9781137353320

Lackey, D. (2019). *How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read – Content For marketers*. Blazon.

Meili, C. (2022). *Create, Connect, and Deceive: Islamic State Supporters' Maintenance of the Virtual Caliphate Through Adaptation and Innovation.* The George Washington University.

Solti I. (2019). Az OSINT információgyűjtő eszközeiről. *Nemzetbiztonsági Szemle*, *7*(2), 3–18. http://doi.org/10.32561/nsz.2019.2.1

Szabadföldi I. (2022). A mesterséges intelligenciával támogatott nyílt információszerzés (OSINT) – evolúció és kihívások. *Nemzetbiztonsági Szemle*, *10*(1), 30–51. https://doi.org/10.32561/nsz.2022.1.3.

Tikk D. (2007). *Szövegbányászat*. Typotex Kft.

Zegart, A. B. (2022). *Spies, lies, and algorithms: the history and future of American intelligence*. Princeton University Press. https://doi.org/10.1515/9780691223087

A cikkben szereplő online hivatkozások

URL1: *Mi az a mesterséges intelligencia?* https://azure.microsoft.com/hu-hu/resources/cloud-computing-dictionary/what-is-artificial-intelligence/#how

URL2: *What is Natural Language Processing?* https://www.ibm.com/cloud/learn/natural-language-processing

URL3: *NoSQL-adatbázis – Mi az a NoSQL?* https://azure.microsoft.com/hu-hu/resources/cloud-computing-dictionary/what-is-nosql-database/

URL4: *Mi az az adat-irányítópult?* https://powerbi.microsoft.com/hu-hu/data-dashboards/

URL5: *6 Reasons Why Open-Source Intelligence is Climbing the Priority Ladder*. https://www.echosec.net/blog/6-reasons-why-open-source-intelligence-is-climbing-the-priority-ladder?utm_campaign=Blog&utm_mediu%E2%80%A6

URL6: *What is OSINT (Open-Source Intelligence).* https://encyclopedia.kaspersky.com/glossary/osint/

URL7: *Exclusive: Inside Hizbollah's fake news training camps sowing instability across the Middle East. The Telegraph*. https://www.telegraph.co.uk/news/2020/08/02/exclusive-inside-hezbollahs-fake-news-training-camps-sowing
URL8: *Drug Enforcement Administration EMOJI Drudg Code.* https://www.dea.gov/sites/default/files/2021-12/Emoji%20Decoded.pdf
URL9: *Kinyerés, átalakítás és betöltés (ETL) – Azure Architecture Center*. https://learn.microsoft.com/hu-hu/azure/architecture/data-guide/relational-data/etl
URL10: *Adattárház összefoglaló*. http://scs.web.elte.hu/Work/DW/adattarhazak.htm
URL11: *List of Public Victims.* https://flashpoint.io/wp-content/uploads/2022/05/Ransomware-Dashboard-1-1-1024x522-2.png

A cikk APA szabály szerinti hivatkozása

Gulyás A. (2023). A nyílt forrásból származó adatgyűjtés automatizálásának lehetőségei. *Belügyi Szemle*, *71*(7), 1237–1269. https://doi.org/10.38146/BSZ.2023.7.6