MIT NYÚJTHAT A MODERN INFORMATIKA AZ IRODALOMTUDOMÁNY SZÁMÁRA?

Mészáros Tamás

PhD, adjunktus,
BME Villamosmérnöki és Informatikai Kar
Méréstechnika és Információs Rendszerek Tanszék
meszaros@mit.bme.hu

Az irodalomtudomány és az informatika sok évtizedes közös múltra tekint vissza, a két terület folyamatos egymásra hatása kölcsönösen új módszerek bevezetését és új tudományos eredmények elérését eredményezte. A cikk röviden áttekinti az informatika robbanásszerű fejlődését, és rávilágít, hogy az elmúlt években teremtett új lehetőségeinek jobb kihasználása érdekében kutatás-módszertani újításokra van szükség az irodalomtudományban. Ezen a területen ugyanis a rendelkezésre álló adatok mennyisége nem nő olyan dinamikus mértékben, mint a biológiában, a csillagászatban vagy a fizikában, így a számítógépek megnövekedett teljesítményében rejlő lehetőségek nem mennyiségi, hanem minőségi javulás útján aknázhatók ki: a feldolgozott szövegek gazdagítása és a „kutatói tudás" gépre vitele révén.

Történeti áttekintés

A számítógépek első alkalmazásai között volt a konkordancialisták készítése (Busa, 1980), a szerzőségi kérdések vizsgálata (Mosteller – Wallace, 1963) és statisztikai elemzések végrehajtása természetes nyelvű szövegeken (Clement, 2014). Később a strukturált dokumentumformátumok alkalmazásában (Burnard – Rahtz, 2002), könyvek millióit tartalmazó digitális könyvtárak létrehozásában (Fox – Sornil, 2003), majd ezekre épülő webes megjelenítő és keresőszolgáltatásokban jártak élen az irodalom- és könyvtártudományi fejlesztések.

A hazai kutatások is több évtizedes múltra tekintenek vissza, számos sikeres digitalizálási munka zajlott le (Bartók et al., 2006), amelyek a TEI XML formátum alkalmazására is kiterjedtek (lásd Palkó Gábor cikke, 1316. o.). A Magyar Elektronikus Könyvtár úttörő kezdeményezés volt az elektronikus dokumentumgyűjtemények létrehozásában (Zimányi, 2001), nyelvészeti eszközök jelentek meg kifejezetten régi irodalmi szövegek elemzésére (Novák et al., 2013), digitális írói szótárak készültek (Kiss, 2012; Mártonfi, 2014), és számos további példát sorolhatnánk a terület eredményei között.

A számítógépek teljesítményének növekedése

Az elmúlt évtizedekben a számítástechnika mind eszközeit, mind módszereit tekintve dinamikus fejlődésen ment keresztül. A rendszerek tárolási kapacitása és az adatok elérési

sebessége évtizedenként nagyjából egy nagyságrenddel lett nagyobb, a kezdeti néhány megabájtos (egymillió bájtos) kapacitású, rendkívül drága rendszereket mára megfizethető, terabájtos (egybillió bájtos) eszközök váltották fel. Míg az IBM által készített első merevlemez nagyjából egy könyvet volt képes tárolni, ma már egy átlagos méretű háttértáron elfér egymillió könyv teljes szövege.

Az informatikai eszközök adatfeldolgozó képessége ennél is nagyobb mértékben növekedett: a teljesítmény/ár arányuk nagyjából három-négyévente egy nagyságrenddel lett kedvezőbb. A kezdetben kevesek által elérhető, drágán üzemeltethető, szekrény(sor) méretű rendszereket előbb a személyi, majd az elmúlt évtizedben sok alkalmazásban mobil számítógépek váltották fel, illetve egészítették ki. Azt a feladatot, amit az ötvenes években még egy hónapig számolt az IBM 704 számítógép, ma egy másodperc alatt képes megoldani egy mobiltelefon. Egy átlagos mai személyi számítógépen a teljes Mikes-életmű konkordancialistájának elkészítése másodpercek kérdése (bár az eredmény diszkre írása egy átlagos merevlemezen egy-két percig is eltarthat).

A mai informatikai rendszerek mind számító-, mind tárolórendszereiket tekintve bőséges, javarészt kihasználatlan kapacitással rendelkeznek az irodalomtudományi kutatások jelenlegi igényeit tekintve. Mindez lehetőséget kínál a mostaninál lényegesen nagyobb erőforrásigényű kutatási módszertanok kidolgozására és alkalmazására is.

Módszertani változások az informatikai rendszerekben

A számítógépes rendszerek kapacitásának növekedése kibővítette az alkalmazható algoritmusok körét is, és az egyre összetettebb funkciókat nyújtó rendszerek újszerű üzemeltetési megoldások kialakítását is ösztönözték.

Az informatikai rendszerek feladatmegoldó képessége, az alkalmazott módszerek sokat fejlődtek az elmúlt évtizedekben. A szövegekkel végzett tevékenységeikre koncentrálva megállapítható, hogy a kezdetben csak ügyes írógépként (szövegszerkesztők) és tárolóeszközként (dokumentumtárak) használt számítógépek egyre inkább az információfeldolgozás és -elérés alapvető eszközévé váltak, és előtérbe került az emberi tudás reprezentálására és felhasználására is képes rendszerek megvalósítása.

Az egyre nagyobb méretű elektronikus szövegtárakban a könyvtártudomány eredményeire támaszkodó elemző- és keresőrendszerek jelentek meg, amelyek a szövegek különféle statisztikai jellemzőit felhasználva végzik el feladataikat. Ennek széles körben ismert és használt példája a web világhálója és keresőrendszerei, amelyek létezésük bő két évtizede alatt teljesen átalakították mindennapjaink információbeszerző tevékenységét. Az átvett módszerek önmagukban azonban nem bizonyultak elégségesnek a felhasználók által elvárt funkciók megfelelő minőségű megvalósításához, így új megoldások kidolgozására volt szükség. Egyre nyilvánvalóbbá vált, hogy a számítógépek akkor tudnak az emberek hasznosabb segítőivé válni ezen a téren, ha az általuk kezelt szövegekről több tudással rendelkeznek, képesek azokat a statisztikai jellemzőiken túlmutató módon feldolgozni és értelmezni.

A kilencvenes években a web szabványait gondozó World Wide Web Konzorcium (W3C) új irányvonalat dolgozott ki a felmerült problémák kezelésére: a szemantikus web koncepciót (Berners-Lee et al., 2001). Ennek alapját az XML strukturált dokumentumfor-

mátum képezi, amely lehetővé teszi szövegek tartalomrészeinek egyszerű jelölését és programozott feldolgozását. Az XML, illetve a hozzá kapcsolódó technológiák és szoftverek (szerkesztő-, feldolgozó és tárolóprogramok) mára széles körben elterjedtek, és számos területen kiegészítették, esetenként fel is váltották a korábbi rendszereket. Az irodalomtudomány az XML első alkalmazói közé tartozik a mára meghatározóvá vált TEI- (Text Encoding Initiative) szabvány kidolgozásával (Burnard – Rahtz, 2002).

A strukturált dokumentumformátumok alkalmazása azonban csak az első lépést jelenti a szövegekben tárolt információk számítógépes kezelése terén. Az informatikai rendszerek viszonya a természetes nyelvű szövegekhez átalakulóban van: egyre több kutatás és gyakorlati megvalósítás foglalkozik a feldolgozáson túlmutató, a szövegek értelmezésére, illetve a bennük tárolt, hozzájuk kapcsolódó tudás gépi reprezentációjára alkalmas rendszerek tervezésével és megvalósításával. Ezek a tudásalapú rendszerek végső soron azt célozzák meg, hogy működésük során a szövegekkel kapcsolatos emberi (szakértői) tudás minél nagyobb részét legyenek képesek tárolni és alkalmazni, olyan felhasználási lehetőségek előtt is megnyissák az utat, amelyeket a korlátos emberi erőforrások miatt eddig nem tartottak reálisnak.

Az informatikai rendszerek használata és üzemeltetése terén is több módszertani változás történt az elmúlt évtizedekben. A nagygépes, centralizáltan menedzselt, kötegelt feldolgozást végző rendszereket felváltották az önállóan üzemeltetett, párhuzamos számításokat is futtató személyi számítógépek és munkacsoportszerverek, majd a számítógépes hálózat megjelenésével kiépültek az ezeket egy rendszerbe integráló hálózati és elosztott rendszerek. Az elmúlt néhány év a virtualizáció és a felhőalapú informatika terén hozott újításokat. Ezekkel a megoldásokkal jelentősen egyszerűbbé válik az erőforrás-gazdálkodás, feladatainkhoz a számítógépes hálózatot és a virtualizált infrastruktúrát felhasználva dinamikusan tudunk fizikai erőforrásokat (processzoridőt, tárolási kapacitást) rendelni.

Ezek a modern, virtualizált és felhőalapú infrastruktúrák ma már a hazai üzleti és akadémiai szférában is elérhetők. Komoly előnyük egyrészt, hogy a nagy beruházási költségű egyedi megoldásoknál jelentősen kisebb költséggel vehetők igénybe (néhány ezer forintos havi díjért egy erős kiépítésű, többprocesszoros virtuális számítógépet bérelhetünk), másrészt a költségeket az aktuálisan igényelt valós számítási kapacitás határozza meg (elkerülhető a felesleges beszerzés és az üresjárat), harmadrészt az így kialakított rendszerek képesek a személyi számítógépeknél és a munkahelyi szervereknél lényegesen nagyobb számítási és tárolási kapacitást is nyújtani. A Sztaki Cloud (URL1), az MTA induló felhőalapú rendszere (URL2) és az NIIF Cloud szolgáltatása (URL3) a hazai kutatói szféra számára kiváló lehetőséget teremtenek az új rendszerüzemeltetési módszerekben rejlő lehetőségek kipróbálására, illetve kihasználására.

Tudásalapú szövegkezelés

A számítógépeinkben és a felhőalapú rendszerekben rendelkezésünkre álló bőséges erőforrásokat tekintve ma már nem az a kérdés, hogy a gépek mennyi idő alatt oldják meg az irodalmi művek feldolgozása során számukra adott feladatokat, hanem sokkal inkább az, hogy képesek vagyunk-e olyan feladatokat adni nekik, amelyek minél teljesebb mértékben kihasználják a bennük rejlő lehetőségeket. Az irodalomtudomány a modern természettudományokkal ellentétben kevésbé adatintenzív, a feldolgozandó szövegek mennyisége nem nő olyan mértékben, így másutt kell keresnünk a számítógépek kiaknázásának lehetőségeit.

Annak érdekében, hogy az informatikai eszközök nagyobb részt tudjanak vállalni a művek feldolgozásában, illetve nehezen kivitelezhető, vagy ma még talán meg sem fogalmazott kutatói feladatokat is meg tudjanak oldani a jövőben, az általuk nyújtott szolgáltatások minőségét és összetettségét kell javítanunk. Mivel a számítógépek addig a mértékig tudnak a segítségünkre lenni, amennyire képesek „megérteni" a tárolt adatokat és az alkalmazott kutatói módszereket, így több tudással kell ellátnunk őket mind a művekre, mind a kutatásokra vonatkozóan, és ezekre épülve új számítási módszereket kell számukra kidolgoznunk.

Az egyik jelentős eredmény ezen a területen a már említett TEI XML formátumhoz kapcsolódik, amely lehetővé teszi a szöveg strukturális felépítésével, illetve egyes tartalmi elemeivel kapcsolatos tudás gépi reprezentációját. Ha egy irodalmi művet egyszerű szövegként tárolunk a számítógépen, akkor nehéz olyan programot írni, amely képes például a benne található földrajzi hivatkozásokat GPS-koordináták formájában meghatározni. Amennyiben azonban TEI-formátumban tároljuk, és a földrajzi hivatkozásokat jelöljük a GEO-címke segítségével, akkor az így tárolt művekből könnyen kinyerhetők a koordináták. Az így bevitt tudás más elemekkel összekapcsolva további módokon is felhasználható a programjainkban.

Az irodalmi művekről nemcsak a művek szövegének TEI-címkézésével, hanem másféleképpen is rögzíthetünk tudást. A művek egészéhez is rendelhetünk adatokat keletkezésükről, szerzőjükről és egyéb tulajdonságaikról. Erre a célra különféle részletezettségű megoldásokkal találkozhatunk (pl. RDFa, Dublin Core, HTML microformats stb.). Az így tárolt adatok egy nagyobb műgyűjtemény feldolgozásakor segíthetnek a keresésben, az időrendi, szerzőségi és más vizsgálatok végrehajtásában. (Ezeket a módszereket egyre szélesebb körben alkalmazzák a web rendszereiben is.) Lehetőségünk van arra is, hogy leírjuk a számítógép számára a művek, részeik és a bennük található személyek, földrajzi helyek stb. viszonyát más entitásokhoz. Ezen a területen például RDF-leírásokat és OWL-ontológiákat alkalmaznak az informatikai rendszerek. Az így létrehozott kapcsolati hálók – különösen egymástól független adatforrások összeolvasztásával – sok érdekes és újszerű vizsgálat alapját képezhetik. Végezetül akár a művekben leírt információkat (tudást) is átalakíthatjuk valamilyen számítógépes reprezentációvá (tudásbázissá). Ezen a téren még sok felfedeznivaló vár a kutatókra.

A történeti és irodalmi szövegek nyelvi elemzése és normalizálása is aktív kutatási terület a számítógépes nyelvészet keretében (pl. Novák et al., 2013). A valószínűségi tudásmodelleket építő és használó elemző- és jelölőrendszerek minősége ma még nem tökéletes, de további tudás bevitelével módszereik javíthatónak tűnnek. A régi szövegek formai sokszínűségéből fakadó bizonytalanságok kibővített írói szótárak létrehozásával és alkalmazásával kezelhetők (Kiss, 2012). Ezek a klasszikus szótári feladatok mellett informatikai rendszerek által megkívánt adatokat és annotációkat is tartalmazhatnak, így növelve azok hatékonyságát. A digitális írói szótárak létrehozása és alkalmazása a szövegek normalizálása és elemzése mellett jelentősen növelheti a szövegtárak keresőrendszereinek

pontosságát, adataik gazdagíthatják a műveket megjelenítő rendszerek szolgáltatásait, és a bennük tárolt tudás önálló elemzések forrása is lehet a jövőben.

A tudásalapú szövegkezelés kihívásai

Az irodalmi művekkel kapcsolatos (emberek által birtokolt) tudás számítógépre vitele a tudásmérnökség feladatkörébe tartozik. Ennek során számos nehézséggel kell szembenézni.

A tudásbevitel egyszerre igényli a szakterület művelőinek (irodalmárok, nyelvészek, történészek) és a tudásalapú rendszerek kialakításában és használatában jártas szakemberek együttműködését. Ráadásul ez utóbbi terület nem tartozik az informatika széles körben ismert és művelt ágai közé.

Az egyes tudásbeviteli módszerek önmagukban is számos problémát hordoznak, és munkaigényük jellemzően lényegesen nagyobb, mint a szinte teljesen automatizálható, statisztikai módszerekkel dolgozó rendszereké. Például egy irodalmi mű elektronikus formára alakítása automatikusan elvégezhető, a karakterfelismerés hibái egy korrektúrázási fázisban könnyen javíthatók. Ezzel ellentétben a TEI XML címkézés kialakítása manuális munka, ráadásul speciális szakértelmet (összetett címkekészlet ismeretét) megkívánó folyamat. Bár az XML-szerkesztő szoftverek számos ellenőrzési lehetőséget kínálnak, a bevitt tudás tartalmi ellenőrzésére sok esetben csak emberi olvasással és értelmezéssel van lehetőség. Hasonlóképpen az irodalmi művekben található hivatkozások RDF-formátumú adatokká alakítása is nehézkes, kevesek által ismert és alkalmazott eljárás.

Ezek a problémák nagyban hátráltatják és sok esetben meg is akadályozzák a tudásalapú szövegkezelés alkalmazását, az így megvalósítható informatikai szolgáltatások kialakítását, az ezekre épülő kutatások elvégzését, így végső soron új tudományos eredmények elérését. Ezen kihívások leküzdése az informatikus és a bölcsész szakemberek közös feladata. Az MTA BTK Irodalomtudományi Intézete és a Budapesti Műszaki és Gazdaságtudományi Egyetem Méréstechnika és Információs Rendszerek Tanszéke által közösen végzett kutatásokban ezen a téren kívánunk eredményeket elérni.

A digitális Mikes kritikai kiadás és a DHmine kutatói rendszer

Az elmúlt években elkészült a Mikes-életmű digitális feldolgozása, valamint az XML-alapú Mikes-szótár (Kiss, 2012), amely a művek teljes szóanyagát tartalmazza (URL4). A következő célkitűzésünk Hopp Lajos kritikai megjegyzéseinek számítógépes feldolgozása, strukturált tárolása és a benne található egyes tudáselemek gépi reprezentációja.

Első lépésként megtörtént a kritikai kiadás digitalizálása és korrektúrázása. A Mikes-művek alapszintű TEI-címkézéséhez képest a kritikai megjegyzéseket egy részletesebb címkézéssel láttuk el, mind strukturális elemeikre, mind a bennük található hivatkozásokra vonatkozóan. A folyamat felgyorsítására egy automatizált XML-címkézőt fejlesztettünk, amely képes volt a kritikai megjegyzések szerkezetének pontos felismerésére és a hivatkozások egy jelentős részének jelölésére is. Jelenleg a hivatkozások (különösen a földrajzi entitások, személynevek és irodalmi művek) felismerésének javításán dolgozunk további elemzési módszerek kidolgozásával.

Az XML-formátumú kritikai kiadás létrehozása után a következő lépés a művek és a kritikai annotációk egységes webes megjelenítésének kidolgozása lesz, majd a jelölt tartalomelemek mint tudásdarabkák tudástárrá szervezése és összekapcsolása más adatforrásokkal (például Sztaki LOD és DBpedia). A már meglévő kritikai megjegyzések digitalizálása és tudástárrá formálása mellett új annotációk készítéséhez is szeretnénk megoldást nyújtani.

A BME által kidolgozott DHmine-rendszer (URL5) nemcsak a digitalizálás és tudástárépítés egyes részfeladatainak végrehajtását tűzi ki célul, hanem kutatócsoportok belső együttműködésének támogatását (fórumrendszerrel, felhőalapú tárhellyel és tartalommegosztással) és terveink szerint a tudományos eredmények adatainak (DataCite-referenciákkal rendelkező) közzétételét is.

Kulcsszavak: *informatika, irodalomtudomány, tudásalapú rendszerek, szövegbányászat, szövegelemzés, XML*

IRODALOM

Bartók István – Golden D. – Horváth I. – Káldos J. – Mayer Gy. – Mártonfi A. – Tóth T. – Vadai I. – Vaskó P. (2006): Digitalizálás, *Magyar Tudomány.* 167, 7, 831–836. • http://www.matud.iif.hu/06jul/09.html

Berners-Lee, Tim – Hendler, J. – Lassila, O. et al. (2001): The Semantic Web. *Scientific American.* 284, 5, 28–37. • http://tinyurl.com/hh3h9m6

Burnard, Lou – Rahtz, Sebastian (2002): The Role of the Text Encoding Initiative (TEI) in the Authoring and Interchange of XML Documents. In: *ELPUB.* • http://elpub.scix.net/data/works/att/02-22.content.pdf

Busa, Roberto (1980): The Annals of Humanities Computing: The Index Thomisticus. *Computers and the Humanities.* 14, 2, 83–90.

Clement, Tanya (2014): Text Analysis, Data Mining, and Visualizations in Literary Scholarship. In: Price, Kenneth M. – Siemens, Ray (eds.): *Literary Studies in the Digital Age.*. Modern Language Association of America • http://tinyurl.com/jmcf7mq

Fox, Edward A. – Sornil, Ohm (2003): Digital Libraries. In: *Encyclopedia of Computer Science*, 576–81. John Wiley and Sons Ltd., Chichester, UK • http://dl.acm.org/citation.cfm?id=1074100.1074337

Kiss Margit (2012): A digitális Mikes-szótár. *Magyar Tudomány.* 173, 3, 279–284. • http://www.matud.iif.hu/2012/03/04.htm

Mártonfi Attila (2014): Számítógép és írói szótár – különös tekintettel a készülő József Attila-szótárra. *MAGYAR NYELV.* 110, 1, 30–46. • http://tinyurl.com/j3y2jhu

Mosteller, Frederick – Wallace, David L. (1963): Inference in an Authorship Problem. *Journal of the American Statistical Association.* 58, 302, 275–309. • https://www.stat.cmu.edu/Exams/mosteller.pdf

Novák Attila – Orosz Gy. – Wenszky N. (2013): Morphological Annotation of Old and Middle Hungarian Corpora. In: *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, 43–48. • http://www.aclweb.org/anthology/W13-2706

Zimányi Magdolna (2001): A Magyar Elektronikus Könyvtár. *Magyar Tudomány.* 2 • http://www.matud.iif.hu/01feb/zimanyi.html

URL1: Sztaki Cloud http://cloud.sztaki.hu/

URL2: MTA http://felho.mta.hu

URL3: NIIF Cloud szolgáltatása https://www.niif.hu

URL4: http://mikesszotar.iti.mta.hu/

URL5: http://dhmine.mit.bme.hu