AZ OKTATÁS, A KUTATÁS ÉS A KÖZGYŰJTEMÉNYEK DIGITÁLIS TRANSZFORMÁCIÓJA FELSŐFOKON

NETWORKSHOP 2024
33. Országos Informatikai Konferencia

2024. április 3–5.
Eszterházy Károly Katolikus Egyetem, Eger

AZ OKTATÁS, A KUTATÁS ÉS A KÖZGYŰJTEMÉNYEK DIGITÁLIS TRANSZFORMÁCIÓJA FELSŐFOKON

NETWORKSHOP 2024
33. Országos Informatikai Konferencia

2024. április 3–5.
Eszterházy Károly Katolikus Egyetem, Eger

Szerkesztette: Tick József, Kokas Károly, Holl András

HUNGARNET Egyesület
Budapest, 2024

HUN-REN
Magyar Kutatási Hálózat

NETWORKSHOP

Szerkesztette: Tick József, Kokas Károly, Holl András

Tipográfia és tördelés: Vas Viktória
Korrektúra: Danyi Melinda
Angol nyelvi lektor: Cseresnyés Dorottya

Networkshop 2024 konferencia előadásainak közleményei
Eszterházy Károly Egyetem, Eger
2024. április 3–5.

ISBN 978-615-82243-2-1
DOI: https://doi.org/10.31915/NWS.2024

Kiadja a HUNGARNET Egyesület
az MTA Könyvtár és Információs Központ közreműködésével
Budapest
2024

Borítókép: freepik.com

TARTALOMJEGYZÉK

DOI: 10.31915/NWS.2024.2

NETWORKSHOP

Adatbányászati gyakorlatok repozitóriumra és MTMT-re

Data mining exercises for the repository and MTMT

Holl András
MTA KIK
holl.andras@konyvtar.mta.hu

Andódy Katalin
BME OMIKK
andody.katalin@bme.hu

Absztrakt

Tanulmányunkban megvizsgáljuk a gépi eljárásokkal lekérdezhető információforrások (esetünkben az EPrints nyílt forráskódú szabad repozitórium szoftver, az MTMT, valamint a doi.org) felhasználhatóságát teljes szövegű publikációk, illetve metaadatok bányászatához. A technikai lehetőségek vizsgálata mellett kitérünk egy konkrét feladatra: a Szociológiai Szemle cikkeinek PDF formátumú anyagában, illetve az MTMT-ben és a REAL-ban való programmatikus vizsgálatokra, melyek célja idézéskapcsolatok keresése volt.

Bevezetés

A tudományos közlemények hagyományosan humán felhasználásra készülnek. Megtalálásukat, felhasználásukat, bibliometriai értékelésüket elősegítendő metaadatokat alkalmazunk. Bár a teljes szöveges keresés a népszerű keresőmotorok elterjedése óta fontos szerepet játszott, a közlemények gépi feldolgozása sokáig a háttérbe szorult. Ebben szerepe van a humán olvasó számára optimalizált formátumoknak, a hibákat tartalmazó OCR-ezett szöveges rétegeknek is.

Az MTA Könyvtára részvételével folytatott, A **Tudomány a Magyar Nyelvért Nemzeti Program részeként megvalósuló A magyar nyelv digitális támogatása a magyar tudományosság szolgálatában** alprogram (Holl et al., 2023) a REAL-ban található szövegek feldolgozását célozza. A kutatási program mellékszálaként idézéskapcsolatok gépi felderítésének lehetőségét vizsgáltuk egyszerű gépi szövegfeldolgozó módszerek alkalmazásával, nagy nyelvmodellek használata nélkül. Abból a feltételezésből indultunk ki, hogy amennyiben egy folyóirat CrossRef-es DOI azonosítókat használ, ezeket az irodalomjegyzékből egyszerű mintaillesztéssel ki lehet nyerni. Mi több, a DOI-k alapján az MTMT-ben egyszerűen és pontosan azonosíthatóak a közlemények. Bár jelentős manuális munkát igényelt, hogy meggyőződjünk a gépi eszközök megfelelő működéséről, és kezeljük a DOI-k hibás alkalmazásból adódó problémákat, a célt C-shell szkriptek és Linux-os környezetben elérhető szabad szoftvereszközök alkalmazásával kívántuk elérni.

I. Kitekintés

Alapvető szerepe van a hivatkozásoknak a bibliometriában, kutatásértékelésben – ezt általánosan elfogadott ténynek tekintjük és a továbbiakban nem foglalkozunk vele. Kiindulási pontunk az, hogy a bölcsészet- és humán tudományok terén nagy jelentősége lenne az MTMT-ben hiányosan reprezentált hivatkozáskapcsolatok teljesebbé tételének.

A hivatkozások szövegbányászatával foglalkozott Kostoff et al. (2001) és Váradi et al. (2014). Ebben az időben a szövegbányászat mintaillesztéses, szabály alapú módszerekkel történt. A nagy nyelvmodellek használatát megelőző szövegbányászati technológiákat foglalja össze Thakur és Kumar (2021). A témának a Scientometrics külön tematikus szekciót szentelt (Cabanac et al., 2020). A témában született külföldi irodalom nem a hivatkozások szövegben történő azonosítására koncentrál, ezen túllépve például a hivatkozás szövegkörnyezetét elemzik. Csak Váradi Tamásék cikke különbözik: a MATRICA projekt közvetlen előzménye volt a a jelenlegi TMNP projekt ezen célkitűzésének.

Mindennek az oka az, hogy a természet-, élet- és műszaki tudományokban, angol nyelvterületen pedig a humán és társadalomtudományokban is, a hivatkozáskapcsolatok jelentős része nagy adatbázisokban rendelkezésre áll. A hivatkozások szövegbányászattal való felderítésére nem lesz már szükség a DOI-azonosítók használatának széles körű elterjedése után sem. Ennek a technikának most (az elmúlt évtizedekre és a közeljövőre) és itt (hazánkban és az angol nyelvterületen kívül) van létjogosultsága.

II. Adatbázisok és alkalmazásprogramozási felületek

A projektben az EPrints repozitóriumi szoftver, az MTMT valamint a DOI Alapítvány alkalmazásprogramozási felületeit használtuk.

EPrints
Az MTA KIK repozitóriuma, a REAL EPrints szoftvert használ.[1] Ez a szoftver igen nagy mértékben nyitott, akár több tucatnyi export-formátum is rendelkezésre áll a teljes szöveg vagy a metaadatok kinyeréséhez (ha a telepítéskor a szükségtelennek ítélt formátumok támogatását ki nem kapcsolják). Az exportok megfelelően szerkesztett URL-el elérhetőek. Mi a JSON kimeneti formátumot használtunk többnyire.

MTMT
Nyitottságra tervezett szoftvert használ a Magyar Tudományos Művek Tára: mind a szerkesztői, mind a nyilvános felület az MTMT API-n keresztül kommunikál a háttérrendszerrel[2].

Az API-hívások URL-jeit megfigyelhetjük a nyilvános felületen <ref>https://m2.mtmt.hu/gui2/</ref> a böngészőnk címsorában.

DOI
A DOI Alapítvány felületét (doi.org) az azonosítók létezésének és a regisztráló ügynökség ellenőrzésére használtuk.

1 https://www.eprints.org/uk/index.php/eprints-software/

2 https://dsd.sztaki.hu/hu/projects/mtmt2

III. Segédprogramok Linux alatt

A legtöbbet használt eszköz a wget volt, bár használhattunk volna curl-t is. További említésre méltó programok a JSON feldolgozásra szolgáló jq[3] és az XML feldolgozásra alkalmas XmlStarlet[4] voltak, mindkettő parancssorból használható, így szkriptekben alkalmazható.

IV. Idézéskapcsolatok keresése

A kutatás résztvevőinek (a kutatók, kutatást végző intézmények és a folyóiratok) hatás-mutatója az idézettség. Ugyanakkor a hivatkozáskapcsolatok a kutatási teret átszövő, kifeszítő hálózatok egyikének élei is. Az MTMT által szolgáltatott fontos információk egyikét a hivatkozások jelentik – majd 12 millió ilyen kapcsolatot tartalmaz az adatbázis. A természet-, élet- és műszaki tudományok területén a nemzetközileg látható hivatkozáskapcsolatokat be lehet gyűjteni nagy kommerciális adatbázisokból. A bölcsész és társadalomtudományok területén ezeket a kapcsolatokat többnyire a szerzőknek kell felfedeznie és felvinnie – így nem meglepő, hogy az adatbázis e téren hiányos. Minden bizonnyal sokkal több idézet van, mint ami az MTMT-ben (vagy a Scopusban, Web of Science-ben) szerepel – de vajon mennyire hiányosak ezek az adatbázisok?

Az idézett közlemény oldaláról ezeket a kapcsolatokat nehéz felderíteni, az idéző közlemények felől könnyebb: csak az irodalomjegyzékeket kell feldolgozni. Azzal a feltevéssel élünk, hogy egy folyóirat több évfolyamában megjelent cikkek összességére a „kimenő" idézetek jellemzői nagyon hasonlóak a „bejövő" idézetekével, azaz a kommunikáció izotróp.

V. Munkamódszer

Egy olyan folyóirattal foglalkoztunk, amelyet a REAL-ban cikkenként archiválnak (így a folyóiratszámok cikkekre bontásának problémája eleve megoldott), és amelyik az MTMT-ben is teljesen feldolgozott: a Szociológiai Szemlével, öt teljes évfolyamra kiterjesztve a vizsgálatot. (A cikkek REÁL-azonosítói az MTMT-ből könnyen kinyerhetőek voltak.) Munkamenetünk a következő lépéseket tartalmazta (ahol szkripteket, API-hívásokat alkalmaztunk, megmutatjuk a lényeges kódrészletet):

1. **Adott évben megjelent (teljes tudományos) folyóiratcikkek lekérdezése MTMT-ből;**
2. **Cikkek külső azonosítóinak, bibliográfiai adatainak letöltése az MTMT-ből táblázatos formában, REAL azonosítók kigyűjtése;**
3. **Cikkek teljes szövegének letöltése a REAL-ból;**

3 https://jqlang.github.io/jq/

4 https://xmlstar.sourceforge.net/doc/UG/xmlstarlet-ug.html#idm47077139529952

A PDF-fájl URI-jának kinyerése REAL-azonosító alapján, JSON-formátumban lekért metaadatokból:

```
wget http://real.mtak.hu/cgi/export/eprint/$item/JSON/REAL-eprint-$item.js
jq -r „.documents[] | .files[] | .uri’ REAL-eprint-$item.js > <URI>
```

A PDF-fájl letöltése wget-tel történik az URI ismeretében:

```
wget --user=******** --password=******** -O <PDF_állomány> <URI>
```

A PDF konvertálása szöveges állományra:

```
pdftotext <PDF_állomány>
```

4. **DOI-azonosítók kigyűjtése a dokumentumból szkript segítségével**

A DOI kinyerése reguláris kifejezés illesztéssel:

```
cat <TXT_filename> | tr -s ‚[[:space:]]’ ‚\n’ | grep ‚10\.[0-9]\{4,9\}/[-._;()/:A-Z0-9]*’|grep -v „[#?]” | sed ‚s/10\./+/’|awk -F’+’ ‚{print „10.”$2}’ | tr ‚\n’ „,’ | sed ‚s/(/%28/g’ | sed ‚s/)/%29/g’ | sed ‚s/;/%3B/g’ | sed ‚s/>/%3E/g’ | sed ‚s/</%3C/g’ | tr -dc ‚[:print:]’|sed ‚s/\.\,/\,/g’
```

5. **talált DOI-azonosítók visszaellenőrzése: doi.org-szolgáltatás segítségével, (létezik-e, illetve milyen DOI ügynökségnél regisztrálták)**

A DOI-k ellenőrzése a doi.org API segítségével (több DOI-t lehet egy hívásban megadni, vesszővel elválasztva, a példában kettőt adtunk meg):

```
wget -O SzocSzem_2018_4.93573.doi https://doi.org/doiRA/10.51624/SzocSzemle.2018.4.8,10.1177/0891243206289499,
```

6. **hibás DOI-azonosítók összevetése PDF-en szereplő DOI-val, DOI-k javítása, hibakeresés;**
7. **javított DOI-listában szereplő azonosítók keresése az MTMT-ben (valójában két lekérdezést végeztünk általában közleményekre és külön csak forrásközleményekre):**

```
wget -O <MTMT_lekerdezes_eredmeny>.js ‚https://m2.mtmt.hu/api/publication?format=json&cond=publicationRole;eq;CORE&cond=identifiers.source;eq;6&cond=identifiers.idValue;eq;<ellenorzendo_DOI>&incFields=mtid’
jq -r „.content[] | .mtid’ <MTMT_lekerdezes_eredmeny>.js
```

8. **felvihető idézéskapcsolatok (manuális) azonosítása.**

VI. Eredmények

Év	(TT) cikkek száma	irodalom-jegyzék tételek száma	Talált DOI-k száma	Ebből MTMT-ben szereplő közlemény	Ebből forrásközl.	Új idézés-kapcsolatok száma
2022	23	1110	442	122	97	28
2021	23	1245	477	96	75	15
2020	26	1078	296	63	52	19
2019	25	883	210	32	21	9
2018	27	1081	205	31	21	7
össz.	124	5397	1630	344	266	78

A folyóirat cikkeiből kinyert azonosítók évfolyam szerint és összesítve (TT: teljes tudományos).

VII. Melléktermékek

Vizsgálatunk során feltártuk a DOI-azonosítók használatának néhány ismétlődő hibáját: gyakran felesleges pont került az azonosító végére, illetve az irodalomjegyzék tételek nem mindegyikénél lett DOI feltüntetve (pedig lehetett volna).

VIII. Lehetséges folytatás

A pilot kiterjesztését tervezzük néhány további folyóiratra, illetve megkíséreljük a jelenleg még manuálisan végzett munkafázisok automatizálását. Meg szeretnénk próbálni az irodalomjegyzékekhez való DOI keresést a CrossRef e célra szolgáló eszközének szkriptelt alkalmazásával – ami nem csak a a kimaradt DOI-azonosítókat találhatja meg, de DOI-t nem alkalmazó folyóiratokra is kiterjeszthetővé teheti az automatizált idézéskapcsolat-keresést. Szükséges lenne a magyar kiadású folyóiratok DOI-prefixeinek összegyűjtésére is.

Összefoglalás

Cikkünkben megvizsgáltuk a hivatkozáskapcsolatok kinyerésének lehetőségét hazai kiadású, magyar nyelvű folyóiratok cikkeiből egyszerű gépi eszközök segítségével. A vizsgálat alátámasztotta feltételezésünket, miszerint szövegbányászattal mindeddig az MTMT-ből hiányzó idézéskapcsolatok deríthetők fel.

A Szociológiai Szemle 124 cikkének feldolgozásával 78 ismeretlen hivatkozáskapcsolatot (MTMT-terminológiával idézéskapcsolatot) találtunk (ebből 46 független). A talált kapcsolatokat fel is vittük az MTMT-be – az általunk felvitt új kapcsolatok az összes 7,4%-át teszik ki. A nagy nyelvi modellek használata a későbbiekben további lehetőségeket tárhat majd fel. Melléktermékként azonosítottuk a DOI-azonosítók használata hazai gyakorlatának néhány hibáját. Tanulságként említhetjük, hogy az itthon is széles körben használt szabad szoftverplatformok és a közösségi alapon működő nemzetközi adatbázisok egyaránt kínálnak lehetőséget a gépi lekérdezésre, adatbányászatra.

Irodalomjegyzék

Cabanac, G., Frommholz, I., Mayr, P.
Scholarly literature mining with information retrieval and natural language processing: Preface. Scientometrics, 125. pp. 2835–2840. (2020) https://doi.org/10.1007/s11192-020-03763-4

Holl András, Prószéky Gábor, Váradi Tamás, Laki László
Repozitóriumi gyűjtemény mint adatkorpusz. Tudományos és Műszaki Tájékoztatás 70 : 2 pp. 164–167. (2023) https://doi.org/10.3311/tmt.13239

Kostoff, R.N. et al.
Citation Mining: Integrating Text Mining and Bibliometrics for Research User Profiling. Journal of the American Society for Information Science and Technology. 52. 13. pp. 1148–1156. (2001)

Thakur, K., Kumar, V.
Application of Text Mining Techniques on Scholarly Research Articles: Methods and Tools. *New Review of Academic Librarianship*, *28.* 3, pp. 279–302. (2021) https://doi.org/10.1080/13614533.2021.1918190

Váradi Tamás, Mittelholcz Iván, Blága Szabolcs, Harmati Sebestyén
Magyar társadalomtudományi citációs adatbázis: A MATRICA projekt eredményei. Magyar Számítógépes Nyelvészeti Konferencia, 10. Szeged. pp. 269–276. (2014) http://acta.bibl.u-szeged.hu/58891/