Lengyelné Molnár Tünde

Kulcsszó-meghatározási technikák

Az információk elérése során a szövegek automatikus összefoglalásának, illetve a kulcsszavak azonosításának alkalmazása egyre jelentősebb szerepet játszik minden területen. A tanulmányban áttekintjük a szöveg-visszakeresés módszereit, a kulcsszavak szerepét a folyamatban, valamint a szövegek visszakeresését szolgáló reprezentánsokat. A tanulmány második felében egy saját kutatás bemutatása történik, melyben nagyszámú minta alapján a kitöltők kulcsszó-megjelölési technikáját elemezzük.

A szöveg visszakeresési lehetőségeit több oldalról közelíthetjük meg. Egyik lehetséges csoportosítása az alábbi hármas tagolódás:[1]

1. Hipertextalapú keresés: heurisztikus keresés, amellyel egy kívánt szövegrészletet keresünk a dokumentum egyes részei-kulcsszavai, illetve a dokumentumok közt kialakított hiperlinkeken keresztülhaladva. A módszer hatékonyságát befolyásolja a linkek kiépítésének minősége, és a karbantartása.
2. Adatbázisban használt lekérdező nyelvén történő keresés: a módszer a strukturált adatok keresésére alkalmas, mely magában foglalja az előnyét: csak azt kell megadni, milyen adatokat akarunk; azt viszont nem, hogyan érjük el őket.
3. A legnépszerűbb a teljes szövegű keresés: alkalmazása során egy kulcsszó segítségével szintaktikai egyezésekre, és nem szemantikai kapcsolatokra keresünk. A módszer gyengesége a szemantikai vizsgálat hiánya.

A kulcsszavakhoz való hozzáállás a 1990-es évek végén esett át egy nagy változáson: az internetes keresőrendszer fejlődésében a 1998-ra eluralkodó spamhullám újra gondolásra késztette a keresőrobot-fejlesztőket. Az addig alkalmazott átolvasása a weboldalaknak, majd a többször előforduló szavak beazonosítása, és a smart indexek általi megfordítása és találatként történő megjelenítését ki kellett egészíteni, fejlesztésre volt szükség.

Bill Gross ötletgyáros a GoTo (a későbbi Overture) internetes keresőrendszer kifejlesztője fogalmazta meg, hogy a keresés lényege a kulcsszóban rejlik. Ha valaki beír egy keresőszót, akkor adatbankot keres, amely a vele kapcsolatos minden információt tartalmaz.[2]

„Minden elhibázott kezdeti lépésem ahhoz a felismeréshez vezetett, hogy a keresés valódi értéke a kereső kifejezésben van… Rájöttem, ha valaki a „Diana hercegnő" kifejezést adja meg a keresőgépnek, végső soron egy olyan Diana hercegnő „üzletben" szeretne kikötni, ahol minden Diana hercegnővel kapcsolatos termék és információ kiterítve hever előtte."[3]

Felismerését továbbfejlesztette és megalkotta a teljesítményalapú modellt: csak azokért a látogatókért kell fizetni, akik a hirdetésre kattintva belépnek az oldalra, és ezt a szolgáltatást az ingyenes keresési lehetőség mellett kínálta a felhasználóknak. A piac azonban akkor még nem volt érett az ötletére, és ez a GoTo vesztéhez is vezetett, ötlete azonban ma a kulcsszóalapú reklámpiac alapmodellje, mely több milliárd dolláros üzlet.

Az információ menedzsmentje és kinyerése egyre fontosabb lett az elmúlt években. Ezért olyan algoritmusokra van szükség, amelyek támogatják, hogy megtaláljuk az óriási mennyiségű információnak azokat a kis darabjait, amelyekre éppen szükségünk van. Az egyik eszközcsoportot erre a feladatra a keresőmotorok jelentik. „A felhasználók alapvető szükségleteinek kielégítésén túl olyan eszközöket is fejlesztenek, amelyek a kifinomultabb igényű felhasználók (közösségek, cégek, érdekcsoportok) számára nyújtanak jóval alaposabban kidolgozott módszereket. Például erre szolgálhatnak a csoportosító vagy osztályozó algoritmusok, illetve más adatbányászati technikák."[4]

Pinto[5] négy csoportra bontja azokat a lehetőségeket melyekkel segíthetjük a dokumentumok, szövegek jellemzését, keresését:

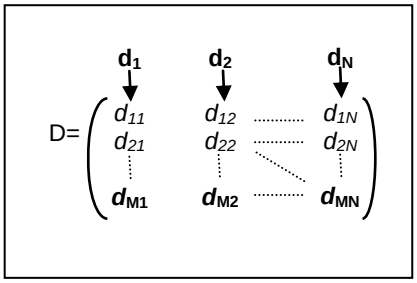
1. Semmivel. A támogatás hiányát úgy érti Pinto, hogy nem kerülhető meg a teljes szöveg visszaolvasása, történjen az géppel, vagy ember által, és a szöveg minden elemét össze kell hasonlítani a keresőkérdéssel.

2. Szabad tárgyszavakkal. Ennek alkalmazása során a dokumentumhoz szabad tárgyszavakat rendelnek hozzá egy adatbázisban, és a szabad tárgyszavakat vetik össze a keresőkérdéssel. Pinto felhívja a figyelmet a homonimák problémájára, és ezek kerülését ajánlja.
3. Deszkriptorok használatával. „Itt már feltételezünk egy hátteret alkotó tezauruszt… …A keresőkérdést is e tezaurusz ellenőrzött terminusainak figyelembevételével fogalmazzuk meg, amivel jelentősen gyorsítható az összehasonlítás folyamata. A szabad tárgyszavakkal kapcsolatos buktatókat így elkerüljük ugyan, de jóval több emberi munkaerőt igényel a háttérben működtetett tezaurusz karbantartása."[6]
4. Tartalmi összefoglalóval. „A tartalmi összefoglalókat jó esetben olyan szakemberek készítik, akik értenek is valamennyire az adott tárgykörhöz. Ez a követelmény ma már egyre képtelenebb a tudomány és technika mai állása mellett. Itt lépnek színre és segítenek a számítógépek és a matematikai nyelvészeti módszerek. A keresés során a teljes dokumentum helyett csak a tartalmi összefoglalóval foglalkozunk, annak szavai, terminusai között keresünk valamilyen módon."[7]

A kulcsszavak szerinti keresés fontosságát jól mutatja *Marwick* 2001-es felmérése[8], mely a vállalati tudásmenedzsment területén mutatta ki, hogy az összefoglalások nélküli (information retrieval) IR-rendszerek használói a keresés során megtalált dokumentumok nagyjából 24%-ának ellenőrzik a relevanciáját, míg az összefoglalásokat tartalmazó IR-rendszereket használók mindössze 3%-a néz utána ennek.[9]

Kulcsszó-meghatározási technikák

1. Vektortérmodell: "A vektortérmodell a szövegbányászati modellek első, klasszikus, erőteljesen a lineáris algebrára építő reprezentációs eszköze."[10] A vektortérmodellben a vektorok értéke az egyedi kifejezések relevanciája, a vektortér dimenziója pedig az egyedi kifejezések száma. A lineáris mátrixban a korpusz dokumentumainak száma határozza meg a mátrix oszlopainak számát (N), a dokumentum egyedi releváns kifejezései pedig a mátrix sorait (M) (*1. ábra*).



1. ábra **Kulcsszómátrix**[11]

„Ez alapján a térbeli struktúra alapján ezután lehetőség nyílik az egyes dokumentumok egymáshoz képesti hasonlóságának feltárására, dokumentumklaszterek definiálására, egyéb jelentéstartalom kinyerésére."[12] Ha túl sok egyedi szót tartalmaz a korpusz, akkor magas a vektortér dimenziós száma, melynek csökkentésére az alábbi módszereket alkalmazhatjuk:

- stopszavazás,
- szótövezés,
- alacsony információtartalmú szavak elhagyása (főelem kiválasztás),
- előbbi kettő inkább az előfeldolgozásban kerül alkalmazásra, míg a harmadik a már előfeldolgozott adatokat alakítja tovább.[13]

A mátrixban szereplő szavak közül a kulcsszó meghatározásához elemzésre kerül, hogy hány dokumentumban milyen gyakori az adott szó, hasonló érték esetén fontos elemezni, hogy hogyan oszlik meg az előfordulás: minden dokumentumban egyenletes az előfordulás, vagy vannak dokumentumok, melyekben koncentráltabban fordul elő az adott szó. Ha vesszük a szóelőfordulások számát (tk), és a dokumentumok számát, melyekben előfordul az adott szó (N), akkor a következő képlettel meghatározhatjuk a vektortérmodell esetén alkalmazott leggyakoribb sémát: idf súlyozást (inverze document frequency):

$$idf\ (t_k)=\log\ (N/n_k)$$[14]

1. Súlyozott gyakoriság (Weighted Term Frequency, WTF) módszere:[15] A kifejező szavak lépésenkénti keresése a súlyozott szógyakoriság. Első lépésként a dokumentumot részekre (bekezdésekre vagy mondatokra) kell bontani. Ezt követően minden szó esetében meg kell határozni a WTF-et és szakaszonként össze-

adni. A szöveg-összefoglaláshoz a legmagasabb összesített értékű szakaszokat kell kivonatolni.

2. WEBSOM-módszer: „A WEBSOM módszer az önszervező térképet (SOM) használja szöveges dokumentumok kétdimenziós térképre való leképezésére. A térképen a hasonló dokumentumok azonos vagy egymáshoz közeli térképelem(ek)en jelennek meg és minden egyes térképelemhez egy mutató is tartozik, ami a dokumentum-adatbázisra mutat. Ezáltal egy keresésnél, miközben azon dokumentumokat megtaláljuk, melyek legjobban illeszkednek a kereső kifejezésre, további releváns eredményeket is találunk, melyek a megtalált dokumentumokat jelképező térképelemmel azonos vagy ahhoz közeli térképelemre voltak leképezve, függetlenül attól, hogy a keresési kifejezésnek megfeleltek-e vagy sem. A WEBSOM-ot kimondottan nagy szöveggyűjteményekben való keresésre dolgozták ki."[16] „…a modell az egységgel kapcsolatos összes dokumentum minden egyes szavához relatív gyakoriságot számol, majd ezeket összehasonlítja a térképen lévő többi egység minden szavának relatív frekvenciájával. A módszer nagyon lassú, és nem praktikus."[17]
3. liGHtSOM-modell: a WEBSOM módosított modellje, mely a súlyok eloszlásán, valamint a beviteli adatok tömörítésére használt random kivetítési mátrix egyszerű módosításán alapul.
4. Katz K-keverék („K-mixture") modell: „A modell egy módosított kifejezés-súlyozás alapján rangsorolja a mondatokat és a magasan rangsoroltakat választja ki a végső összefoglalóhoz. Az ismétlődő mondatokat eltávolítja és egy csempézett összegzést készít."[18]

A módszerek gyakorlati szükségessége megkérdőjelezhetetlen, mivel az internetes keresések 65%-a információkeresésre irányul.[19]

Az a minimális elvárás a kereső személyek részéről, hogy egy-egy kifejezéshez kapcsolódva releváns találatot szeretnének, például egy reklámhoz kapcsolódva rögtön a cég honlapját találják meg, ezért a keresőknek márkára épülő kampányokat közvetlen válaszra épülővé kell átalakítani, a keresőrendszereknek tovább kell lépni a kulcsszó meghatározásán: figyelni és elemezni kell az emberek kulcsszavait, és a találatok közül történő kiválasztási technikáját (clikstream, kattintáskutatások), valamint alkalmazni a nyelvészeti technikák eredményeit. Nézzük meg a legelterjedtebb módszereket!

Nyelvészeti technikák

Mélyszemantikájú indexelés (Latent Semantic Indexing, LSI)

A *mélyszemantikájú indexelés* olyan technika, amely képes a szavak közötti jelentésbeli, szemantikai információkat megragadni, így olyan – egyébként releváns – dokumentumokat is találatként visszaadni, amelyekben az eredeti lekérdezés egyik szava sem fordul elő; tehát képes a szavak közötti látens viszonyokat és szemantikai összefüggéseket modellezni. Ez elsősorban idegen nyelvű szöveg esetén válhat fontossá, amikor a szóegyezésekre aligha támaszkodhatunk. A módszer előnye, hogy *thesaurus* használata helyett a mélyszemantikájú indexelés mindig az aktuális tématerületű korpusz esetében képes automatikusan feltérképezni a szavak közötti jelentésviszonyokat.

A módszer lényege a *szinguláris érték-dekompozíció (singular value decomposition, főkomponens-dekompozíció)* műveletében rejlik, amely hasonló a sajátérték dekompozícióhoz és a faktoranalízisben használt módszerhez. A szinguláris érték-dekompozíció eredménye vektorok egy halmaza, amelyek rendre az egyes egyedi szavak és dokumentumok pozícióját reprezentálják a redukált k dimenziószámú térben. Információ-visszakeresés során a lekérdező *sztring* által adott szavak azonosítanak egy pontot az LSI-térben, azaz, a lekérdezés az általa tartalmazott egyedi szavak helyvektorainak súlyozott vektoriális összege által meghatározott helyen fog szerepelni. Ezt követően a dokumentumok rangsorolása a lekérdezés LSI-térbeli helyzetéhez való közelségük alapján történik, tipikusan koszinusz távolsági mértékkel számítva. Mivel az egyedi szavak és a dokumentumok is ugyanabban a térben helyezkednek el, így lehetőség nyílik azok tetszőleges kombinációjú összehasonlítására, úgymint az egyedi szóhoz legközelebb eső dokumentumok, az egyedi szóhoz legközelebb eső más egyedi szavak, a dokumentumhoz legközelebb eső egyedi szavak és a dokumentumhoz legközelebb eső dokumentumok kimutatására.[20]

Kulcsszóosztályok kivonatolása

A módszer célja meghatározni és jellemezni a témákat nagyméretű szövegekben (10 millió szavas nagyságrend esetén), még pedig úgy, hogy téma szerinti alszövegekre bontjuk az eredetit. Ezek alapján a két fő megoldandó probléma:

- megtalálni a témákat a szövegtesten belül,
- karakterizálni őket és meghatározni az adott témák feltűnési helyét a szövegben, hogy kiválaszthassuk azokat a szövegrészeket, amelyekből az alszövegeket összeállítjuk.

Kulcsszóosztályok kivonatolása esetén a generálni kívánt alszöveget a következő lépés alapján képezzük, amely lexikális forrásokat hoz létre szöveges adatokból. Ennek feltétele, hogy:

- Az alszövegeknek önálló szövegként is meg kell állniuk a helyüket. Ezért mondatok helyett bekezdéseket használ a szövegegység a kivonatoláshoz, de elfogadja, hogy egy bekezdés több témára is tartalmazhat hivatkozást.
- Figyelembe kell venni, hogy a következő lépések hagyatkoznak az eredményre, ezért a kivonatolt alszöveg következetességére nagyobb hangsúlyt kell fektetni, mint a teljességére.
- Az optimális eredmény elérése érdekében nem előre definiált témalistákra hagyatkozik a módszer, hanem hagyja, hogy a szövegből *„kerüljenek elő”* a témák. Ez azt is jelenti, hogy a rendszer szemantikai információkat is ki tud nyerni a vizsgált szövegből.
- A legfontosabb, hogy a folyamat teljesen automatizált legyen, és ne igényeljen emberi beavatkozást, vagy külső adatok betáplálását.

Ezek teljesüléséhez a témákat tematikus kulcsszóosztályok használatával kell kivonatolni és leírni, azaz szavakkal, amelyek jelenléte az adott szövegrészben szorosan összefügg egyes témák felbukkanásával. Ezek a kulcsszóosztályok külső beavatkozás nélkül kerülnek elő a szövegből egy háromlépéses rendszernek köszönhetően:

1. A szöveg kisebb részeiből néhány tökéletlen osztályt vonnak ki egy klasszikus hierarchikus csoportosítási technikával.
2. Ezeket a halmazokat aztán összehasonlítjuk és ütköztetjük, hogy kisebb, de zajmentes halmazok megbízható és következetes jellemzését kapjuk.
3. Az így keletkezett osztályok aztán egy egyszerű, felügyelt tanulási módszer alapját adják, hogy a témákat jobban lefedő csoportokat képezzünk.

A generált kulcsszóosztályok pontosan visszaadják a témákat, amelyek a vizsgált szövegekben előfordulnak; a szöveg tartalmának teljesen komplett és informatív áttekintését teszik lehetővé, így magas precizitással állapíthatjuk meg a témák előfordulásait a szövegben. Mivel a módszer teljesen automatikus, emellett független a forrástól és a szöveg nyelvétől, sokféleképpen használható nagyméretű szövegek feldolgozására: besorolás, indexelés, szűrés, visszakeresés.[21]

Szótövezés

A szótövezés olyan szavak szótőre redukálását jelenti, amelyek valamilyen jelentésmódosító ragot, toldalékot, *prefix*et vagy *suffix*et kaptak. Szövegbányászati szempontból sokszor az ilyen szavak között nem teszünk különbséget. A szótövezés különösen fontos a ragozó nyelvek, így például a magyar nyelv esetében, ahol a ragok vagy egyéb toldalékok az eredeti szóhoz hozzátapadnak. Ekkor ugyanis ugyanannak a szónak igen sok variánsa előfordulhat, amelyeket a szótövezés folyamán mind egy közös őshöz kell visszavezetni. A szótövezés eredményeként a korpuszban figyelembe vett egyedi szavak száma csökken, hiszen adott szóvariánsokat a szótövükkel helyettesítjük. Természetesen a legtöbb elterjedt szótövező algoritmus angol nyelvterületen használatos, a legnépszerűbbek a következők:

- Paice/Husk szótövező algoritmus,
- Porter szótövező algoritmus,
- Lovins szótövező algoritmus,
- Dawson szótövező algoritmusa,
- Krovetz szótövező algoritmusa.[22]

Stopszó-eliminálás

A módszer eredménye, hogy a korpusz már csak a számunkra releváns szavakat fogja tartalmazni. A folyamat során az olyan gyakori, de relevanciával nem rendelkező szavakat töröljük, amelyek általában minden dokumentumban jelen vannak, de nem hordozói a dokumentumspecifikus jelentésnek, ezért csak megnehezítik a tudás kinyerését. Az ilyen szavak tipikus példái:

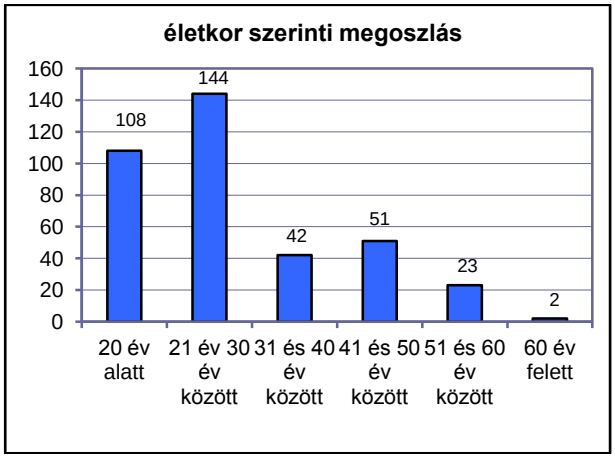
- névelők,
- névutók,
- névmások,
- kötőszavak,
- kérdőszavak.

Szűrésük *stopszólista* segítségével lehetséges; ha az adott szó szerepel a listán, töröljük. Összeállítása során alkalmazhatjuk a TF-IDF-módszert, amely minden szóra megadja korpusz feletti fontossági súlyt. Ezt követően az első N darab legkisebb súlyút átemelhetjük a stopszólistára.[23]

A következőben nézzük meg egy konkrét felmérés eredményét, mely az emberek kulcsszó-meghatározási technikáját vizsgálta!

Kulcsszavak a kulcsszavak?

Az *Eszterházy Károly Főiskolán* futó TÁMOP-4.2.2.C-11/1/KONV-2012-0008 „IKT a tudás és tanulás világában – humán teljesítménytechnológiai (Human Performance Technology) kutatások és képzésfejlesztés" pályázat keretein belül történt felmérés során 500 személyt kerestem meg, és többek közt két szakcikk kulcsszavainak megjelölésére kértük a kitöltőket. A kitöltésben résztvevők két mintacsoportot alkottak: hallgatók és szakemberek. A felmérésben részt vett 375 érvényes kitöltőjének korbeli összetételén láthatjuk, hogy a kitöltők 2/3-a 30 év alatti, de több mint 100 kitöltő a 30 évtől idősebb korosztályból is részt vett a felmérésben (*2. ábra*).



2. ábra **A felmérés résztvevői életkor szerint**

Az online felmérés során a kitöltőket az alábbi két cikk elolvasására és kulcsszavainak megjelölésére kértem:

- *Forgó Sándor*: Az új média és az elektronikus tanulás[24]
- *Komenczi Bertalan*: A digitális pedagógus – elméleti megközelítések, fogalom meghatározások[25]

A felmérés során megjelölt leggyakoribb kulcsszavak hatékonyságát az alábbi módon vizsgáltam:

- Felkértem a két szerzőt, adják meg az általuk kulcsszavaknak tartott kifejezéseket.
- Az általam készített szoftverrel meghatároztam a két cikk szógyakorisági listáját.
- Elkészítette a két cikk szófelhőjét a *http://www.wordle.net/* weboldallal, amely szógyakorisági alapon határozza meg a kulcsszavakat.

Az eredményt két mintacsoportra vonatkoztatva elemeztem, cikkekként.

Forgó Sándor cikke esetén a 10 leggyakoribb kifejezést az 1. táblázat tartalmazza. Látható, hogy a kulcsszavak fele mind a négy elemzési módszernél megtalálható, illetve további két kifejezés megjelölésre került három különböző módszer/csoport esetén. Összességében a 10 legtöbbek által megjelölt kulcsszó közül csupán egy-egy van, amelyet nem jelölt meg a szerző vagy a szoftver.

1. táblázat
A 10 leggyakoribb kifejezés (Forgó)

	Word frequency	Szerző	Könyvtárosok	Hallgatók
1	learning	e-learning	média	e-learning
2	tanulási	interaktív	e-learning	digitalizáció
3	digitális	(új) média	digitalizáció	elektronikus
4	hálózati	tanulás	új	Média
5	alapuló	elektronikus	elektronikus	tanulási
6	közösségi	multimédiás	e-learning2.0	új
7	média	2.0	tanulási	e-learning2.0
8	interaktív	online	interaktív	webkettőn
9	elektronikus	hálózati	digitális	IKT-kompetenciákat
10	kommunikációs	kommunikációs	kommunikáció	interaktív

A kulcsszavak szófelhője is tükrözi a magas fokú egyezést (*3. ábra*):

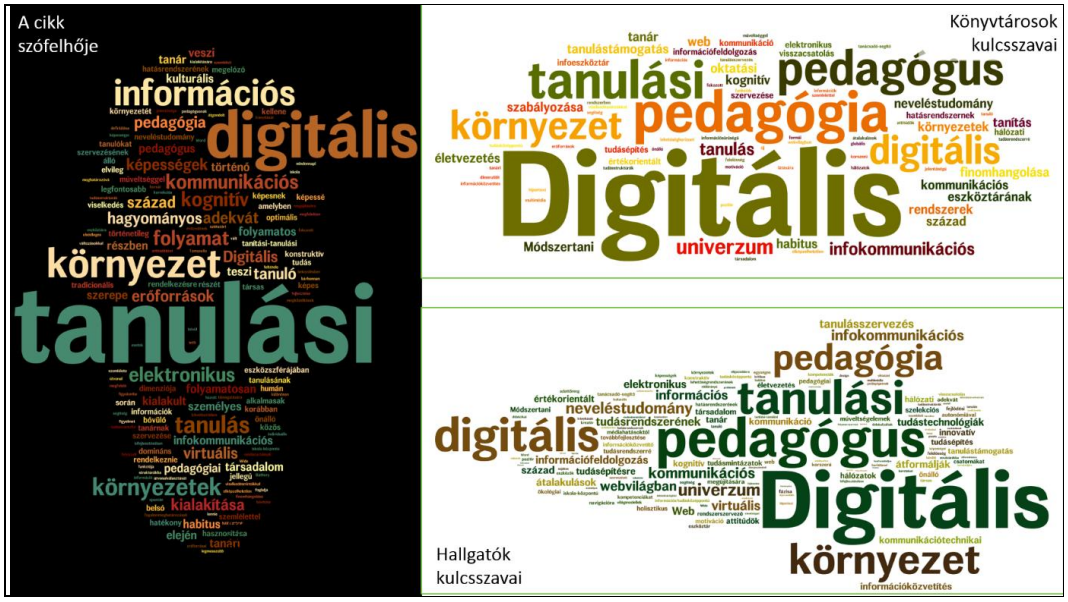
3. ábra **Kulcsszavak szófelhője (Forgó)**

Komenczi Bertalan cikke esetén az egyezés nem ennyire látványos, de a módszer hatékonysága itt is egyértelműen látható. A mintacsoportonkénti kulcsszavak táblázata, amely kiegészítésre került a szógyakorisági lista által generált eredményekkel, illetve a szerző által megjelölt kifejezésekkel (*megjegyzés: nem kulcsszavakat, hanem kifejezéseket jelölt meg a szerző*), a tíz leggyakoribb kulcsszó közül három található meg mindegyik mintacsoportnál, azonban további 4-6 olyan kulcsszó található meg két mintacsoportnál.

2. táblázat
A 10 leggyakoribb kifejezés (Komenczi)

	Word frequency	Szerző	Könyvtárosok	Hallgatók
1	learning	kognitív habitus	digitális	digitális
2	tanulási	tanulási környezetek	pedagógia	tanulási
3	digitális	virtuális dimenziója	tanulási	környezet
4	környezet	didaktikai design	környezet	pedagógus
5	információs	kompetencia	pedagógus	pedagógia
6	tanulás	digitális pedagógia	univerzum	univerzum
7	kognitív	mezovilág	infokommunikációs	neveléstudomány
8	elektronikus	személyes tanulási hálózat	tanulás	infokommunikációs
9	folyamat	kommunikációs hatásrendszer	kognitív	Web
10	kommunikációs	hatásrendszere	habitus	kommunikációs

Az eredeti cikk szófelhője, és az összes megjelölt kulcsszóból készült szófelhő hasonló eredményt tükröz:



4. ábra **Komenczi Bertalan: A digitális pedagógus – elméleti megközelítések, fogalommeghatározások**

Konklúzió

A kulcsszó-meghatározásnak jelentős matematikai háttere van, de a hálózatok és a hálózaton elérhető információk növekedése, a hatalmas méretű korpuszok elérhetősége a téma további fejlődését fogja maga után vonni. A saját felmérés is azt támasztja alá, van értelme kutatni a területet, a kulcsszavak behatárolhatóságát, azonosíthatóságát támasztják alá a humán kutatások is.

Hivatkozások

1 ABRAHAM, Kiryo: Business Intelligence. Aufgaben, Prozess und Architektur [elektronikus dokumentum]. München, 2008. p. 42−43.

2 TÓTH Erzsébet: Hatékony információkeresés a weben. – Nyíregyháza, Örökségünk könyvkiadó, 2010.

3 SHUEN, Amy: Die Web-2.0-Strategie. Köln, 2008. XI, 38. p. Ford: Mizera Tamás

4 CARAMIA, Massimiliano – FELICI, Giovanni: Mining relevant information on the Web. A clique-based approach. = International Journal of Production Research, 14 (2006), p. 2771.

5 FÜREDI Mihály (ford.): Metainformációk előállítása. = Tudományos és Műszaki Tájékoztatás, 51. évf. 12. sz., 2004. [PINTO, Maria nyomán: Engineering the production of meta-information: the abstracting concern. = Journal of Information Science, 29. Vol., 5. No., 2003, p. 405–417.] Elérhetőség: *http://tmt.omikk.bme.hu/show_news.html?id=3781&issue_id=457* [2014.04.12.]

6 lásd előző

7 lásd előző

8 MARWICK, A. D.: Knowledge Managment Technologie. = IBM System Journal 40 (2001) 4. p. 824.

9 NOHR, Holger: Grundlagen der automatischen Indexierung. Ein Lehrbuch. Berlin, 2003. p. 107.

10 Szimulációs környezet. 2010. p. 4. [Elektronikus dokumentum] *http://palyazat.webstar.hu/gop/servlet/download?type=doc_field_file&field=file&id=4669*

11 Szövegbányászat /szerk. TIKK Domonkos. – Budapest: Typotex, 2006. p. 32.

12 Szimulációs környezet. 2010. p. 4. [Elektronikus dokumentum] *http://palyazat.webstar.hu/gop/servlet/download?type=doc_field_file&field=file&id=4669*

13 lásd előző

14 Szövegbányászat /szerk. TIKK Domonkos. – Budapest: Typotex, 2006. p. 36.

15 PRIBE, Torsten – KOLTER, Jan – KISS, Christine: *Semiautomatische Annotation von Textdokumenten mit semantischen Metadaten*. = Wirtschaftsinformatik, 2005. eEconomy, eGovernment, eSociety. Heidelberg, 2005. p. 1319.

[16] ALTRICHTER Márta − HORVÁTH Gábor − PATAKI Béla − STRAUSZ György − TAKÁCS Gábor − VALYON József: Neurális hálózatok. *http://www.tankonyvtar.hu/en/tartalom/tamop425/0026_neuralis_4_4/ch10s03.html*

[17] lásd előző

[18] SARAVANAN, M. – RAMAN, S. – RAVINDRAN, B.: A probablistic approach to multi-document summarization for generating a tiled summary. = International Journal of Computational Intelligence & Applications, 2 (2006). pp. 231−243.

[19] TÓTH Erzsébet: Hatékony információkeresés a weben. – Nyíregyháza, Örökségünk könyvkiadó, 2010.

[20] VÁZSONYI Miklós: Mélyszemantikájú indexelés [elektronikus dokumentum]. Copyright 2006. = URL *http://www.vazsonyi.hu/szovegbanyaszat/14.html* (letöltés: 2011. 10. 30.)

[21] ROSSIGNOL, Mathias – S´EBILLOT, Pascale: Combining statistical data analysis techniques to extract topical keyword classes from corpora. = Intelligent Data Analysis 9 (2005), pp. 105−127.

[22] VÁZSONYI Miklós: Szótövezés [elektronikus dokumentum]. Copyright 2006. = URL *http://www.vazsonyi.hu/szovegbanyaszat/6.html* (letöltés: 2011. 11. 03.)

[23] VÁZSONYI Miklós: Stopszó eliminálás [elektronikus dokumentum]. Copyright 2006. = URL *http://www.vazsonyi.hu/szovegbanyaszat/7.html* (letöltés: 2011. 11. 02.)

[24] FORGÓ Sándor: Az új média és az elektronikus tanulás. = Új pedagógiai szemle, 2009. (59. évf.) 8−9. sz. p. 91−96.

[25] KOMENCZI Bertalan: A digitális pedagógus - elméleti megközelítések, fogalommeghatározások. = LÉVAI Dóra − TÓTH-MÓZER Szilvia − SZEKSZÁRDI Júlia (szerk.): Digitalis_de_generacio 2.0. Budapest: Underground Kiadó és Terjesztő KFT, 2013. p. 193−202.

Beérkezett: 2014. IX. 5-én.

Lengyelné Molnár Tünde
az Eszterházy Károly Főiskola Humáninformatika Tanszékének tanszékvezetője, főiskolai docens.
E-mail: *mtunde@ektf.hu*

Nyílt forráskódú böngészőt ad ki ingyen az Ericsson

Az OpenWebRTC rugalmas, platformfüggetlen WebRTC klienskeretrendszer, amely mind natív WebRTC alkalmazások, mind a böngésző hátterét biztosító rendszerek (back-end) kiépítésére alkalmas.

Az Ericsson Research bejelentette, hogy ingyenesen, nyílt forráskóddal adja ki a *Bowser* elnevezésű webböngészőt és az alapjául szolgáló OpenWebRTC keretrendszert. Az *Ericsson* több választási lehetőséget és nagyobb rugalmasságot szeretne biztosítani a fejlesztők számára, ezáltal felgyorsítva az innovációt a WebRTC közösségben. A WebRTC rendkívül egyszerű módot kínál valós idejű hang-, video- és adatátviteli alkalmazások fejlesztésére. A *World Wide Web Consortium* (W3C) és az *Internet Engineering Task Force* (IETF) szervezetekben szabványosítás alatt álló API-kból és protokollokból áll.

Az OpenWebRTC arra a meggyőződésre épül, hogy a WebRTC szabvány túllép a szokásos böngészőkörnyezeten, és a natív alkalmazások is a WebRTC ökoszisztéma fontos részévé válnak, ugyanazokat a protokollokat és API-kat használva. Ez különösen a mobilplatformok esetében fontos, ahol a natív alkalmazásokat gyakran előnyben részesítik a webalkalmazásokkal szemben. *Stefan Ålund*, az *Ericsson Research* kutatási menedzsere kijelentette: „Mióta 2012-ben a nyilvánosság elé álltunk a Bowserrel, számtalan kérés érkezett hozzánk, hogy osszuk meg fejlesztésünket. Most nemcsak a Bowsert adjuk ki, hanem az alapjául szolgáló, platformfüggetlen WebRTC keretrendszert is, amelyet az Ericsson Research-nél fejlesztettünk ki és már több éve használunk”.

Az Ericsson Research tekintélyes részt vállal a WebRTC szabványosításában: a kezdetektől fogva a szabvány több prototípus megvalósítását fejlesztette ki. Mind az IETF, mind a W3C szabványosítás megkövetel legalább két független, együttműködő megvalósítást. Alund így folytatta: „A WebRTC szabvány folyamatosan fejlődik. A fejlesztők újabb és újabb módokat találnak a technológia mindennapos használatára. Mérnökeink úgy építették ki az OpenWebRTC-t, hogy rendkívül egyszerű legyen módosítani és bővíteni, teret hagyva az API-kkal és az új funkciókkal folytatott további kísérletezésnek”. A Bowser nemcsak nyílt forráskóddal jelent meg, de az Apple App Store-ból is ingyenesen letölthető lesz.

/Forrás: *http://sg.hu/cikkek/108364/nyilt-forraskodu-bongeszot-ad-ki-ingyen-az-ericsson/*

(B.Bné)