Tichy-Rács Ádám

Kutatási projektek hasonlóság szerinti rangsorolása – szemantikus szolgáltatás a Nemzeti Kutatás-nyilvántartási Rendszerben

A Nemzeti Kutatás-nyilvántartási Rendszer újszerű szemantikus alkalmazást hozott létre, amelynek alapján értékelhető két kutatás-fejlesztési projekt hasonlóságának mértéke. A cikk a közismert Boole-algebrai keresési eljárásokkal összehasonlítva, ábrákkal és magyarázatokkal mutatja be az eljárás algoritmusát. Az alkalmazás hozzáférhető, kipróbálható az NKR nyilvános felületén.

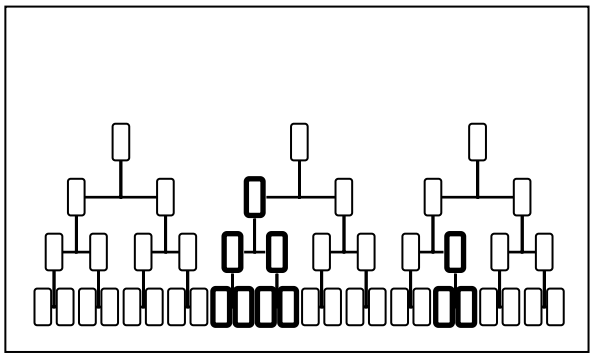
A *Nemzeti Kutatás-nyilvántartási Rendszer (NKR)* a közpénzből finanszírozott kutatás-fejlesztési projektek adatait tartja nyilván 2002 januárja óta. A nyilvántartás megfelel az *Európai Unió* által ajánlott *CERIF (Common European Research Information Format)* szerkezetének. A projektek rövid leírása mellett a tartalom jellemzésére az *Európai Bizottság* által, a projektjavaslatok elektronikus benyújtását támogató rendszer *(EPSS = Electronic Proposal Submission System)* tezauruszát alkalmazza, ami a tezaurusz fejlesztésének egy korábbi fázisában *Ortelius-tezaurusz* néven vált ismertté. A tezaurusz a kutatók által megadott kulcskifejezésekkel folyamatosan bővül magyar és angol nyelven. [1]

A tezaurusz szemantikai struktúrája lehetővé teszi, hogy megkeressünk minden olyan projektet, amely a tezauruszból kiválasztott viszonylag általános kifejezésekből Boole-algebrai módszerrel összeállított kifejezésnek *megfelel*. A keresési eljárás logikai értelemben nem tesz különbséget a találati halmaz elemei között, így azok megjelenítési sorrendje tipikusan a keresőrendszertől különálló komponensben dől el. Ennek eredménye egy olyan sorrend lehet, amelyet a keresés logikájához képest mellékes szempontok határoznak meg: a projektcím betűrendje, a projekt kezdési vagy befejezési időpontja, a támogatás összege. A Boole-algebrai eljárás nem mutatja meg, hogy melyik projekt *felel meg legjobban* a keresési feltételeknek, illetve melyik projekt *hasonlít legjobban* egy előzőleg kiválasztott projekthez. Az NKR legújabb fejlesztése eredményeként meg tudjuk határozni az egyes találatok *relevanciáját* (projektalapú keresésnél: a *hasonlóság mértékét*), így lehetővé vált a projektek rendezése a keresési kifejezés szempontjából lényeges jellemzőjük szerint. A továbbiakban a keresőkérdéssel induló, és a projekthasonlóságon alapuló kereséseknél egyaránt a találatok relevanciájáról beszélünk.

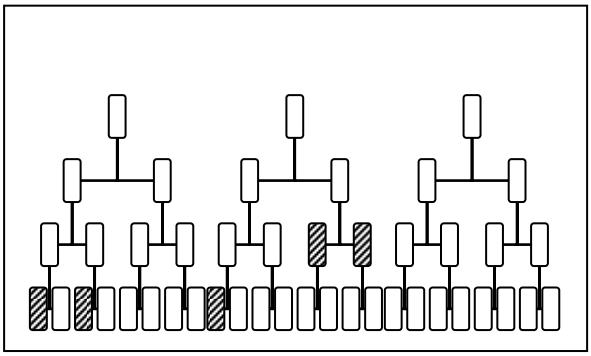
Boole-algebrai keresés tezaurusz segítségével

A jobb áttekintés érdekében ismételjük át a tezaurusszal támogatott Boole-algebrai keresés logikáját. Az egyszerűség kedvéért az alábbiakban csak a VAGY kapcsolattal felépített kifejezést mutatjuk be. Az ilyen típusú keresőkérdés lényegében így fordítható le: keressük mindazokat a projekteket, amelyeket a felsorolt kifejezések, vagy azok tezaurusz szerinti alárendeltjeinek bármelyike jellemez.

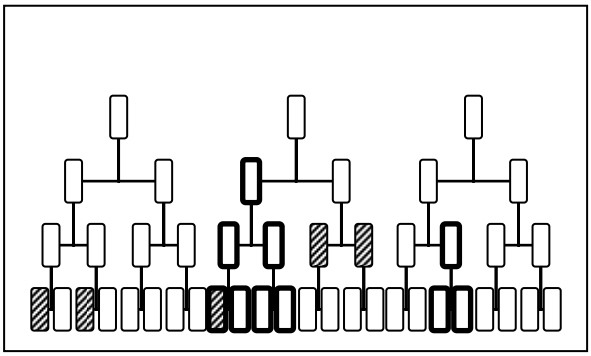
A keresés vizuálisan, egy egyszerűsített, három legfelső szintű elemet tartalmazó bináris tezauruszban leírt keresőkérdéssel mutatható be (*1. ábra*). Az egyes projekteket ugyanebben a tezauruszban írjuk le (*2., 3. ábra*).



1. ábra **A keresőkérdés elemei és azok alárendeltjei (A keresőkérdéshez tartozó kifejezéseket jelképező mezők kerete vastagított)**

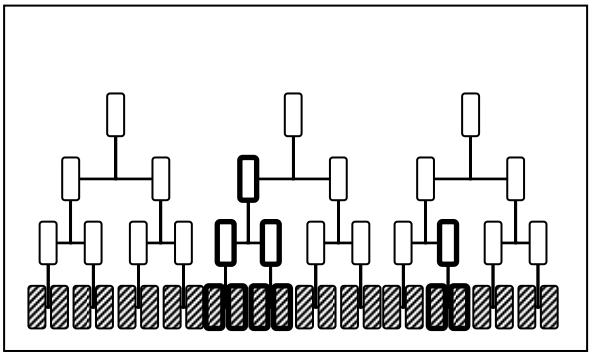


2. ábra **Egy projektet jellemző kifejezések elhelyezkedése a tezauruszban (A projekthez tartozó kifejezéseket jelképező mezők vonalkázottak)**



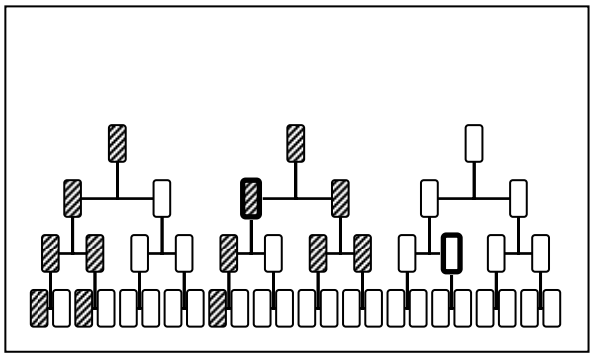
3. ábra **A fenti projekt (2. ábra) eleme a keresőkérdés (1. ábra) találati halmazának, mert az alsó szint egyik eleme megfelel a keresési feltételnek (A közös mezőket a kétféle kiemelés együtt mutatja)**

A Boole-algebrai keresés egyik hátránya, hogy arra projektek „optimalizálhatók", ha a tezaurusz legalsó szintjének valamennyi elemét hozzájuk rendeljük. Nyilvánvaló, hogy az így preparált projekt belekerül bármely – kizárást nem tartalmazó – keresőkérdés találati halmazába (*4. ábra*).



4. ábra **A keresésre „optimalizált" projekt minden keresőkérdés találati halmazába bekerül**

A keresési eljárás leírása megfordítható. Ebben a reprezentációban a keresőkérdés csak a felhasználó által kiválasztott kifejezéseket tartalmazza, és a projekthez rendeljük a kifejezések összes fölérendeltjét. A korábbi keresési példa (1. ábra keresőkérdése és a 2. ábra projektje) a fordított reprezentációban a következő ábrával jellemezhető (*5. ábra*). A kétféle reprezentáció – a találati halmazokat tekintve – egyenértékű. Az utóbbi esetben vagy tárolni – és természetesen a tezaurusz minden módosításával aktualizálni – kell a fölérendeltek listáját és maga a keresés nagyon könnyen végrehajtható, vagy a fölérendelteket a keresés közben rendeljük a projektekhez; ekkor a keresés végrehajtásához szükséges számítási erőforrás lesz nagyobb.

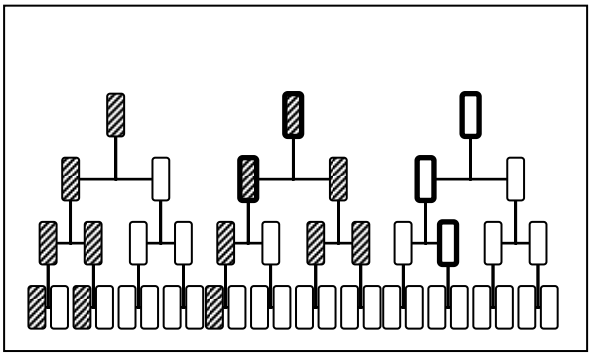


5. ábra **A fenti keresés (3. ábra) a keresőkérdés és a projekt reprezentációjának megfordításával**

Akárhogyan hajtjuk végre e keresést, a fenti projekt (2. ábra) és az optimalizált projekt (4. ábra) ugyanúgy része lesz a találati halmaznak, és, amint azt a bevezetőben láttuk, megjelenítésük sorrendjét tipikusan a projekt és a keresőkérdés közötti relevanciához (l. a következő szakaszban) képest mellékes szempontok határozzák meg.

A relevancia értelmezése

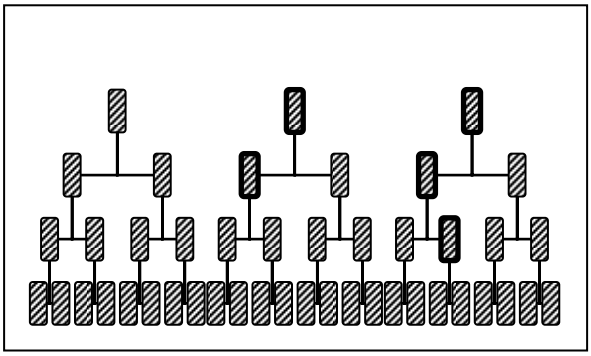
A relevancia mértékének maghatározására az NKR-ben bevezetett mérőszám a keresőkérdés és a projekt kifejezéseinek és azok fölérendeltjeinek korrelációja, vagyis a közös elemek száma osztva a két kifejezésben külön-külön szereplő kifejezések számának mértani közepével. A „tezauruszra vetített képek" alapján a korábbi példában bemutatott keresőkérdés és projekt közötti megfelelés a 6. ábrán látható.



6. ábra **A fentebb bemutatott keresésnek (3. ábra, 5. ábra) megfelelő relevancia meghatározásának sémája**

A bemutatott esetben a projekthez rendelt kifejezések száma 13, a keresőkérdéshez rendelt elemek száma 5, a közös elemek száma 2, amint az jól látható. A számított relevancia (korreláció) mértéke $R=2/(13*5)^{½}\approx0,25$.

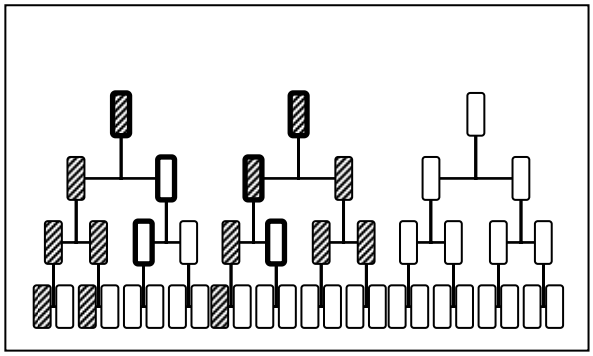
A korábbi, keresésre optimalizált projekt (4. ábra) relevanciája ugyanezen keresőkérdésre: *(7. ábra)* $R=5/(45*5)^{½}\approx0,33$, de relevanciája a tezaurusz méretének növelésével csökken. Az EPSS tezauruszánál 0,04, míg az NKR keretében épülő tezaurusz esetén <0,02.



7. ábra **Keresésre optimalizált projekt relevanciájának meghatározása**

Relevanciaalapú keresés tezaurusz segítségével

Érdemes felhívni a figyelmet arra, hogy a fenti értelemben bizonyos relevanciája annak a projektnek is van a keresőkérdésre vonatkoztatva, amit nem találunk meg, ha a fenti bemutatott reprezentációval végezzük a Boole-algebrai keresést. Ilyen eset fordul elő, ha túlzottan precíz keresőkérdést használunk, vagyis a keresésben felsorolt kifejezések a tezaurusz alacsonyabb szintjén találhatók, mint a projektre jellemző kifejezések, vagy a tezaurusz másik ágán helyezkednek el (*8. ábra*).



8. ábra **Keresőkérdés relevanciája a bemutatott reprezentáció szerinti Boole-algebrai kifejezéssel nem kiválasztott projektre is lehet 0-tól különböző: $R=3/(13*6)^{½}\approx0,34$**

Mint az jól látható, a bemutatott példaprojekt (2. ábra) az új keresőkérdésre sokkal relevánsabb találat (R≈0,34), mint az a korábbira (R≈0,25) volt.

Míg a fent bemutatott reprezentáció szerint végzett Boole-algebrai keresés nem, az NKR új keresőrendszere megtalálja az összes, nem nullarelevanciájú projektet valamely keresőkérdésre.

A relevanciaalapú keresés megvalósításához meg kell változtatni a korábbi, a keresőkérdés és a projekt szempontjából aszimmetrikus reprezentációt: a keresőkérdés és a projekt esetében egyaránt a jellemző kifejezések és azok összes fölérendeltjével kell dolgoznunk, ami ily módon lesz konform a bevezetett relevanciaértelmezéssel.

A korábbi példákhoz (3., 5. ábra) képest szembeötlő a projekt és a keresőkérdés alkalmazásának szimmetriája az eljárás során, ami felhasználható két projekt S hasonlóságának értelmezésére is. A hasonlóság meghatározásakor az egyik projekt leírását tekintjük keresőkérdésnek, és így számítjuk a relevanciát az egyes projektek és a keresőkérdés (itt: szintén projekt) között a korábban megismert módon, azaz a jellemző kifejezések és azok fölérendeltjei közötti korreláció kiszámításával.

Relevancia szerinti sorrendezés megvalósítása az NKR-ben

Az NKR keresőfelületén a szoftver legújabb fejlesztésének eredményeként megjelent beállítási lehetőség, hogy a Boole-algebrai keresés (a felüle-

ten: egyszerű keresés) mellett választható a bemutatott relevancia szerinti keresés és a kettő kombinációja is. Ez utóbbi esetben csak azokat a projekteket rendezi a szoftver relevancia szerint, amelyek a Boole-algebrai keresési feltételeknek (l. 1. szakasz) is megfelelnek.

A relevancia szerinti keresés nemcsak az NKR keretében ülő 19 000 elemű tezaurusszal valósítható meg, hanem a szorosabban vett, 2073 elemű Ortelius-tezaurusszal és a tudományágak és tudományterületek mindössze kétszintű, 63 elemű listájával is. Lehetőség van arra, hogy a relevanciát a projektekben közreműködő szervezetekre, illetve a projekteket megvalósító személyekre értelmezzük – ez utóbbi esetben az adatbázisban megjelenő hierarchiáról nem érdemes beszélni.

Az NKR felületén, a projekt címe mellett egy nem túl feltűnő ikon (*9. ábra*) kínálja azt a lehetőséget, hogy az adott projektből automatikusan generálódjon a lehető legpontosabb keresőkérdés, és azt a rendszer olyan módon futtatja le újabb beavatkozás nélkül, hogy a kutatók által megadott kifejezések alapján épülő tezaurusz segítségével előállítja az adatbázisban tárolt összes többi, legalább minimális hasonlóságot mutató projekt rangsorát.

Elektromágneses és szeizmikus események kapcsolata

9. ábra **Projektcím és mellette a projektek hasonlóság szerinti rangsorolását kiváltó ikon**

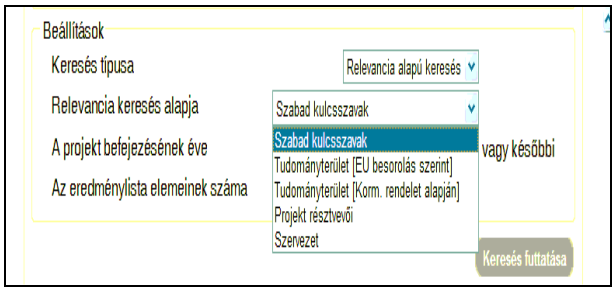
Az ikonra kattintva a következő eredmény adódott (*10. ábra*):

Eredmény: 1-10/6196

Projekt neve	Relevancia
Elektromágneses és szeizmikus események kapcsolata	100%
Elektromágneses jelek terjedése a Föld légkörében	81%
Trimpi kutatás	52%
SAS2-K2 repülőpéldány	52%
SAS2-P1 repülőpéldány	52%
SAS2-P1-TM	52%
Lokális földrengések teljes hullámforma inverziója	50%
Napfénytartam és globálsugárzás interpolációs módszereinek továbbfejlesztése	50%
Távérzékelésen alapuló párolgásszámító algoritmus	50%
Földi elektromágnesség	48%

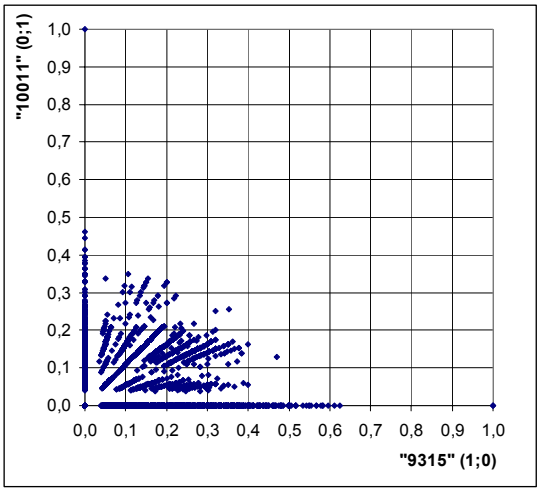
10. ábra **A fentebb bemutatott (8. ábra) projekthez hasonló projektek listájának első tíz eleme**

Kényelmi okokból a rendszer alapértelmezetten nem mutatja magát a keresőkérdést, mert a felhasználó esetleg csak hosszas görgetéssel jutna el az eredménylistáig. Ugyanakkor megtekinthető, a keresésben szereplő kifejezések módosíthatók, a hasonlóság megítélésének szempontjai könnyen megváltoztathatók (11. ábra).



11. ábra **Hasonlóságvizsgálati szempontok módosítása**

Az eljárást két különböző, egymással ortogonális – 0 hasonlóságú, vagyis egyetlen közös kifejezést sem tartalmazó – projekttel egyidejűleg végrehajtva meghatározható az összes többi projekt „helye” a két (10011 és 9315 számú) projekt által kifeszített tudástérben (*12. ábra*). Több viszonyítási projekttel végrehajtott vagy többféle szempontú hasonlóság[1] egyidejű meghatározását az NKR webes alkalmazása nem támogatja, arra csak szolgálati felületen van lehetőség.



12. ábra **Az NKR-ben tárolt projektekhasonlósága két ortogonális projekthez képest**

[1] Ilyen feladat a projektek halmazának két különböző tezaurusz szerinti leírásának összehasonlítása, vagy a projekt tartalmi és a megvalósítók szervezeti hasonlóságának összehasonlítása.

Összefoglalás

A hasonlóság szerinti rendezési eljárással az NKR újszerű szemantikus szolgáltatást kínál. Az eljárás nem túlságosan számításigényes, így alkalmazása javasolható nagyobb információs rendszereken is.

A bemutatott eljárást a NKR munkatársai dolgozták ki. Az alkalmazáshoz szükséges fejlesztéseket az NKR szoftverén az IQSYS végezte el.

Az NKR a *https://nkr.info.omikk.bme.hu* címen elérhető, szabadon, ingyenesen, regisztráció nélkül használható.

Irodalom

[1] TICHY-RÁCS Ádám: A Nemzeti Kutatás-nyilvántartási Rendszer a BME OMIKK-ban. = TMT, 51. köt. 1. sz. 2004. p. 3–15.

Beérkezett: 2011. I. 3-án.

Tichy-Rács Ádám
fizikus, fizikatanár, kulturális menedzser,
2001-től a Nemzeti Kutatásnyilvántartási rendszer vezetője a BME OMIKK-ban.
E-mail: *atichy-racs@omikk.bme.hu*

Jó és nem ajánlott online vírusirtók

Az *Inter Storm Center* (*ISC*) arra figyelmeztet, hogy vannak olyan vírusírók, akik igyekeznek kihasználni a biztonsági oldalak hiányosságait. Emellett előfordul az is, hogy egyes internetes vírusirtó szolgáltatások együttműködnek a kártevők készítőivel. Az ISC szerint, aki egy online vírusirtóval szeretne átvizsgáltatni egy gyanús fájlt, az jobban teszi, ha körültekintően választ, különben kínos meglepetés érheti. A szervezet segíteni szeretne a felhasználóknak, ezért felsorolt számtalan jónak minősülő, kockázatos, illetve kifejezetten nem ajánlott online vírusirtó szolgáltatást.

A világhálón keresztül elérhető biztonsági csomagok egy részének az az előnye, hogy több vírusirtó motorját is felhasználják, így az internetező gyorsan megállapíthatja, hogy vajon egy téves riasztásról van-e szó, vagy valóban kártevőt tartalmaz egyik fájlja. Az ISC összesen öt olyan portált sorolt fel, amelyek használatát nyugodt szívvel ajánlja, ezeket a honlapokat zöld jelzéssel emelték ki. Köztük van a *Virustotal.com*, a *filterbit.com*, a *virscan.org*, a *scanner.novirusthanks.*org és a *virusscan.jotti.org.*

A második kategóriába a sárga jelzésű szolgáltatások tartoznak, az ISC ide négy oldalt sorolt. Ezek többségéről nem állapítható meg, hogy teljesen biztonságosak. Az utolsó, piros jelzésű kategóriába hat honlap került. Ezek mindegyikét vagy korábban összefüggésbe hozták kártevők terjesztésével, vagy még jelenleg is ezzel gyanúsítják. A szolgáltatások ártalmatlan doménneveket használnak, ám azt mindenesetre fontos megjegyezni, hogy a jó online vírusirtók rendkívül hasznosak lehetnek és akkor is segíthetnek, amikor a számítógépre telepített társuk már nem képes megtisztítani a kártevőktől a PC-t vagy a notebookot.

/SG.hu Hírlevél, 2011. február 7., *http:www.sg.hu*/

(SzP)