DOI: 10.31915/NWS.2019.10

Egészségügyi informatikai adatbázisban való szöveges keresés mesterséges intelligenciával támogatott szemantikus keresővel

Kovács Béla Lóránt
Neumann Technology Kft.
bela.lorant.kovacs@negentropics.com

Text search in a health database with a semantic search engine supported by artificial intelligence My purpose is to present a search engine which can process large documents- even books- in the query and search in the medical database based on them. The process works without metadata, so it is enough to upload the search text to the database once, the application executes all the rest of the operations. Searching methods are supported by algorithms based on semantical modeling of natural languages, working in any language. Namely, the algorithm learns languages from the databases. However, the search engine is not only capable of conducting semantic searches but define topics by texts that can be monitored automatically by the program. This process is supported by an artificial intelligence what is suitable for determining semantical groups in one hand, on the other hand, it is learning from the database expansion and user behavior – based on which it can improve the hits, - at the same time, it does not collect any data about the user himself, thus complying with the strictest European data protection rules. I intend to present my description through practical examples demonstrating the innovations that make it unique. At the end, I would like to show in detail what measurable performance the software has and how it differs from other search engines.

Keywords: semantic search, semantic classification, artificial intelligence, search in text documents

A Neumann Technology Kft. a Gazdaságfejlesztési és Innovációs Operatív Program keretén belül, az Európai Regionális Fejlesztési Alapból és hazai központi költségvetési előirányzatból támogatott GINOP-2.1.7-15-2016-00069 azonosító számú projektben egy unikális szoftver prototípusán kezdett dolgozni. A pályázat eredményeként olyan programot fejlesztettünk, amely webes felületen elérhető és általa a felhasználó képes orvosi szövegekben keresni úgy, hogy az eddig rendelkezésre álló keresőknél pontosabb találatot kapjon.[1] A termék a következő elemekből áll:

- Előfeldolgozó
- Index adatbázis
- Kereső adatbázis
- Kereső modul

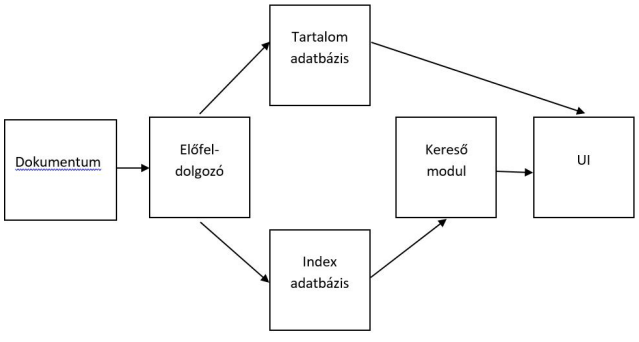
1 Méréseinket 2019 márciusában végeztük. Ennek során kifejezetten az orvosi cikkeket tartalmazó adatbázisok és a hozzájuk kapcsolódó keresők vizsgálatára koncentráltunk, az olyan vizsgálatoktól, amely webes keresők pontosságát is mérte volna, mint amilyen Bennett Shenkeré, tartózkodtunk. (v. ö.: [1])

- Végfelhasználók által elérhető kereső funkciókhoz kialakított felület
- Karbantartó munkatársak által használható funkciókhoz kialakított felület
- Vezető munkatársak által használható funkciókhoz kialakított felület

A keresőmotor, amelyet az Aleetheia névre kereszteltünk, két szempontból is figyelemre méltó: egyrészt a természetes nyelvek szemantikájának modellezését és a mesterséges intelligencia alkalmazását a korábbiaktól eltérő módon oldja meg, másrészt a korábbi, hasonló keresőkhöz képest pontosabb és teljesebb találati listát ad. Éppen ezért a továbbiakban a program rövid bemutatása után ez utóbbi tulajdonságára összpontosítok a terméknek – minthogy a felhasználók számára is ez a legfontosabb[2] – és azt kívánom bemutatni, hogy mit értünk az alatt, hogy pontosabb és teljesebb találati listája, mint a más szoftvereké.

A program működése

A felhasználók a publikus interfész rétegen keresztül kapcsolódhatnak a szolgáltatási réteghez, amelynek a segítségével az alkalmazás funkciókat érhetik el, például a tartalom szerinti keresést. A szolgáltatási rétegben a vezérléseket megfelelő szolgáltatási csatornán fogadjuk, azokat értelmezzük és ütemezetten végrehajtjuk. A szolgáltatási réteg képes fogadni a parancssori és a webes szolgáltatási kéréseket, és az elemző ezeket ütemezetten fogadva és végrehajtva kapja meg. Az alkalmazás magjával készítjük a dokumentum előfeldolgozását. Különböző formátumú szöveges dokumentumok konvertálása, a konvertált tartalmak átadását végzi az elemző részére. Az előfeldolgozó képes a különböző dokumentum formátumok konvertálására. Elvégezzük a tárolást, mely modul feladata kettős: egyrészről tárolni az előfeldolgozó által átalakítandó bemeneti forrásokat, illetve az elemző által kategorizált kimeneti adatokat, döntésekhez szükséges információkat. Tárolja a tanuló adatbázist, a számított eredményeket, illetve megfelelő formátumra konvertált dokumentumokat. Ezt követően elemzéseket végzünk, az előfeldolgozott forrásokon a kategorizáláshoz szükséges döntéseket hozó algoritmusok végrehajtásával (tanulás, osztályozás, döntés, klaszterezés), illetve a feldolgozott források meghatározott struktúrában való elhelyezésével. Másrészt az említetteken kívül az instrukciók nélküli, tartalom alapján végzett csoportképzéshez szükséges algoritmusokat is tartalmazza és futtatja.

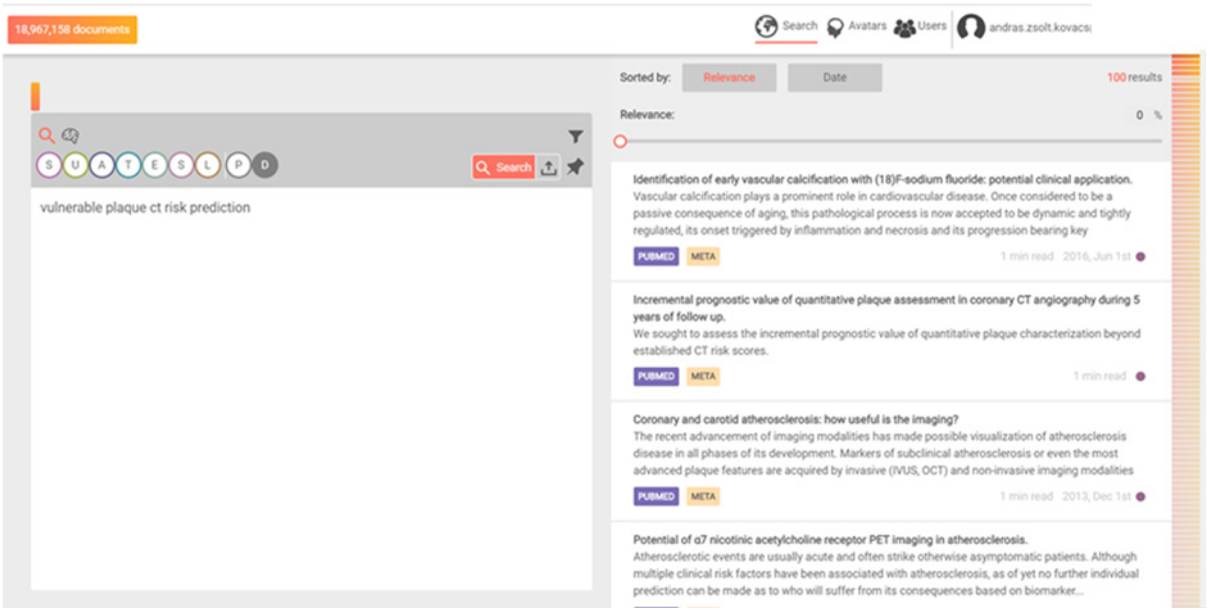


1. ábra: A program felépítése

2 A felhasználók érdeklődéséről: a találatok pontosságáról és teljességéről bővebben lásd.: [2]

A program pontossága

A pontosságot úgy mértük, hogy kutató orvosokat és doktoranduszokat kértünk meg arra, hogy a PubMed adatbázisán egy-egy keresés kapcsán osztályozzák a találatok pontosságát.[3] A találatokat egytől tízig pontozták úgy, hogy a tízes találat ért a legtöbbet, az egyes pedig a legkevesebbet, és minden keresésnél a találati lista első húsz elemét vizsgálták meg. A saját találatainkat két keresési algoritmussal adtuk meg: az első kulcsszavak, a második teljes szövegekkel történő keresésre alkalmas. Ezeket a találatokat vetettük össze a PubMed saját keresőjével, amely ugyan alapvetően kulcsszavas, de metaadatokat is figyelembe vesz.[4] A mi algoritmusaink nem vettek figyelembe metaadatokat és a mérés során nem használtunk keresésre olyan hosszúságú szövegeket, amelyek a PubMed keresőjén nem futnak le. Arra is figyeltünk, hogy ne hibridizáljuk a különböző algoritmusaink által kínált eredményeket, mert ugyan ezáltal még nagyobb pontosságra tehettünk volna szert, ugyanakkor elfedtük volna az egyes algoritmusok valódi képességeit. A továbbiakban éppen ezért a PubMed 19 millió dokumentumot tartalmazó adatbázisán lefuttatott keresések kapcsán három találati listát hasonlítunk össze: a saját kulcsszavas, a saját teljes szöveges algoritmusaink, illetve a Pubmed keresője által adottakat. A kereső felület a következő módon néz ki:



2. ábra Kereső felület (Jobb oldalon a találati lista, bal oldalon a kereső ablak az algoritmusválasztóval)

A felületen látható az algoritmusválasztó, amely a termék végleges változatában nem lesz a felületre kivezetve, mivel a választást az emberek helyet hibridizációs algoritmusok végzik. Számos további funkcióval rendelkezik a program, amelyeket

3 A méréseket 2019. januárja és márciusa között végeztük négy orvos és két doktorandusz bevonásával.

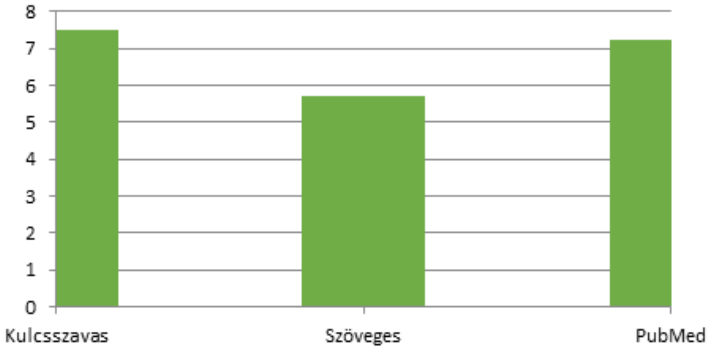
4 A hagyományos kulcsszavas keresők esetében – amilyenek például az egészségügyi adatbázisokban vagy a vállalati adatbázisok keresőiben működnek – kétféle eljárást szoktak kombinálni: az első a keresett kulcsszó gyakoriságából, a második a hozzá kapcsolódó metaadatok (tárgyszavak, tag-ek) előfordulásából szokott kiindulni. (v. ö.: [4].)

azonban hely hiányában nem tudunk bemutatni (témákhoz tartozó szófelhők és szóstatisztikák, trendek, szerzőkre vonatkozó statisztikák stb.).

Első mérés

A keresőbe írt kifejezés a következő volt: *„vulnerable plaque ct risk prediction".*

Az Aleetheia kulcsszavas algoritmusa 75%-os pontosságot ért el, a szöveges azonban csak 57%-ot. A hagyományos kereső 72%-os pontosságra volt képes.



1. Az első keresés pontossága

A találatok részletes elemzéséből kiderül, hogy a kulcsszavas kereső annak ellenére adott pontosabb találatokat, hogy nem használt metaadatokat. Ráadásul nem csak pontosabb, hanem teljesebb is volt a listája, mint a hagyományos keresőnek, hiszen olyan szövegeket is megtalált, amelyeket a másik nem. A szöveges kereső ugyanakkor rosszabb eredményeket adott, köztük teljesen hibásakat is. Ezzel szemben a PubMed keresője ugyan helyes találatokat adott, ám mindössze tizenkét darabot.

Második mérés

A keresőbe írt kifejezés a következő volt: *„cardiovascular risk prediction scores".*

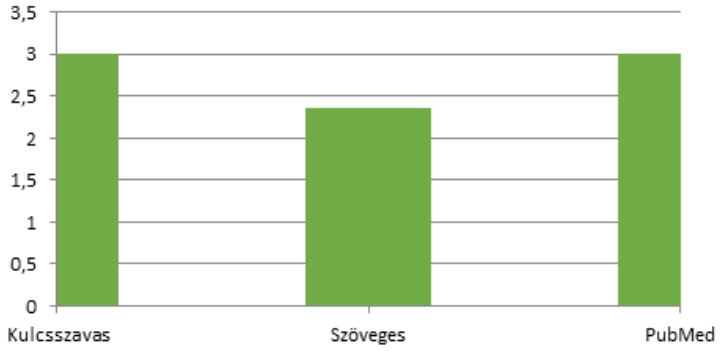
Az Aleetheia kulcsszavas algoritmusa itt is 75%-os eredményt ért el, ám itt már a szöveges kereső is 73%-os pontosságra volt képes. A PubMed itt azonban már csak 64%-os teljesítményt mutatott.

2. A második keresés pontossága

Az eredmények alaposabb elemzéséből az derül ki, hogy az Aleetheia kulcsszavas és a szöveges keresője is jó eredményeket hozott, ám a kétféle algoritmus nem ugyanazokat a jó találatokat jelenítette meg – bár metszete főleg a kiemelkedően jó találatok esetében volt a két listának. Ebből arra a következtetésre jutottunk, hogy érdemes a kétféle keresés erényeit egyesíteni. A PubMed keresőjének kudarcát teljessége magyarázza. A keresett kulcsszavak sok dokumentumban felbukkantak és mivel a PubMed keresőjének természetes nyelvi szemantikus képességei nem voltak, így számos hibás találat is bekerült a listába. Itt látszik, hogy a teljesség növekedése hogyan megy a pontosság kárára a hagyományos keresők esetében.

Harmadik mérés
A keresőbe írt kifejezés a következő volt: *„microvesicles microparticles cardiovascular plaque".*

Ezt a találati eredményt azért mutatom be, mert nagyon jól szemlélteti, hogy mik a különbségek a hagyományos kulcsszavas és a természetes nyelvek szemantikáján alapuló keresők között. Ha csak a listák pontosságát nézzük, akkor szigorú értelemben az Aleetheia kulcsszavas keresője ugyanolyan eredményre volt képes, mint a PubMed-é, szerény 30%-os pontosságra, míg a szöveges kereső ennél is rosszabbra, 24%-ra.

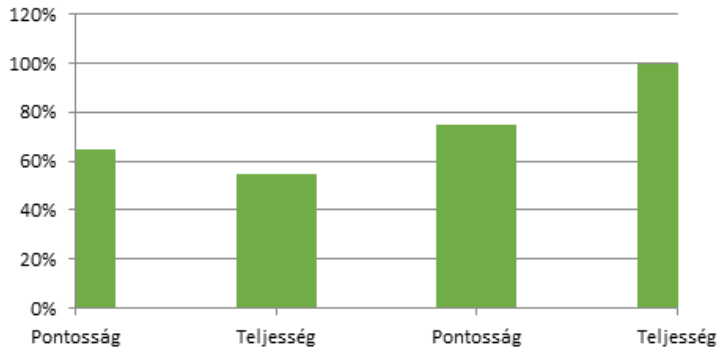


3. A harmadik keresés pontossága

Az eredmények alaposabb tanulsága azonban lényegesen árnyalja a képet. A kulcsszavas keresők ugyanis csak azokat a találatokat jelenítik meg, amelyek minden kulcsszót tartalmazzák vagy pedig minden kulcsszóhoz tartalmazó metaadattal kapcsolatban vannak. Éppen ezért egyetlen, alacsony pontszámmal rendelkező szöveget jelenítenek meg. A szöveges azonban nem ebből indult ki, így számos értékelhetetlen dokumentum mellett több olyat is adott, amelyeket a kutatók sokkal jobbaknak találtak, mint azt, amit a kulcsszavasak adtak.

Tanulságok

Méréseink során több száz keresési listát elemeztünk és mindenhol hasonló eredményre jutottunk – eltekintve természetesen néhány mérési hibától és meglepő szélsőségtől. Az Aleetheia algoritmusai 70-75%-os pontosságra voltak képesek100%-os teljesség mellett, míg méréseink szerint a hagyományos motorok, köztük a PubMed keresője 60-65%-os pontosságra 55-60%-os teljesség mellett.[5]



4. ábra Kékkel a PubMed, narancssárgával az Aleetheia keresőjének pontossága és teljessége

Az algoritmusok hibridizációjával azonban a pontosság lényegesen növelhető volt 90% közelébe. Mindezt a költséges és időigényes metaadatolás nélkül sikerült elérnünk. Keresőnk különösen jó eredményeket volt képes elérni azokban az esetekben, amelyekben mások nem tudtak. Ilyenek a hosszú szöveges keresések vagy pedig azok a kulcsszavas keresések, ahol a kereső kifejezés csak néhányszor fordult elő, így olyan dokumentumokat is érdemes volt átnézni, amelyek szemantikus kapcsolatban állnak a keresett szóval, de nem tartalmazták azt. Ez a kereséstípus meglepő és szokatlanul pontos eredményeket képes elérni a legnagyobb teljesség mellett.

Összegzés

Az Aleetheia a szemantikus keresés új lehetőségeit teremtette meg. Metaadatok használata nélkül is képes a korábbiaknál pontosabb találatokat adni. A termék hordozza azon előnyöket, melyeket terveztünk – azaz nyelvfüggetlenül működik, mobil eszközön is megjeleníthető, gyors, és pontos keresési találatokat ad. Az orvosi alkalmazás kifejezetten jó teszteredményeket hozott, és tudományos-szakmai területen a PubMed adatbázison orvos-kutatóink továbbra is használhatják a saját kutatási területük újdonságainak megjelenítéséhez.

5 Az eredmény fontosságát mi sem bizonyítja jobban, mint az, hogy más keresők, így a Google Scholar vagy a MEDLINE keresője sem képes ilyen eredményre (v. ö.: [5]

Irodalomjegyzék

[1] Bennett S Shenker, The accuracy of Internet search engines to predict diagnoses from symptoms can be assessed with a validated scoring system. International Journal of Medical Informatics, 83(2), 131-139, February 2014 doi: 10.1016/j.ijmedinf.2013.11.002

[2] Ovidiu Dan, Brian D. Davison, Measuring and Predicting Search Engine Users' Satisfaction, Journal ACM Computing Surveys (CSUR) Volume 49 Issue 1, July 2016 Article No. 18 doi: 10.1145/2893486.

[3] Udo Kruschwitz, Charlie Hull, "Searching the Enterprise", Foundations and Trends in Information Retrieval: 2017, Vol. 11: No. 1, pp 1-142. doi: 10.1561/1500000053

[4] Iqra Safder, Saeed-Ul Hassan, DS4A: Deep Search System for Algorithms from Full-Text Scholarly Big Data 2018 IEEE International Conference on Data Mining Workshops (ICDMW) doi: 10.1109/ICDMW.2018.00186

[5] Bramer, W.M., Giustini, D. & Kramer, B.M.R. Comparing the coverage, recall, and precision of searches for 120 systematic reviews in Embase, MEDLINE, and Google Scholar: a prospective study. Syst Rev 5, 39 (2016) doi: 10.1186/s13643-016-0215-7