Automatizált tartalmi feltárás – teszt egy szakirodalmi adatbázisban

Napjainkban a metaadatok és a teljes szöveggel elérhető dokumentumok számának ugrásszerű növekedése miatt a tartalmi feltárás megkönnyítése érdekében előtérbe kerültek az automatikus indexelési eljárások. Ilyen módszereket a média – és sajtódokumentáció már régebb óta sikerrel alkalmaz, pénzt és időt takarítva meg.

Könyvtári körökben az 1990-es években az online katalógusokra való átállással került napvilágra az automatizált tartalmi feltárás kérdése. Német nyelvterületen a számítógép segítségével végzett feltárással kapcsolatos projektek – MILOS I és II, KASKADE, OSIRIS és PETRUS – nyelvészeti és statisztikai eljárásokat vettek alapul.

Nagyjából tíz éve szakinformációs központok is foglalkoznak a problémával. Példa erre a *Pszichológiai Információs és Dokumentációs Központ (Zentrum für Psychologische Information und Dokumentation, ZPID)* által épített, számítógépes nyelvészeten alapuló indexelő szoftvert alkalmazó PSYNDEX. A szoftver a dokumentum címe, absztraktja és a szerzők által adott tárgyszavak alapján javasol deszkriptorokat az intellektuális feltárás elősegítésére.

Erőforrások megtakarítása érdekében a társadalomtudományokra szakosodott *GESIS – Leibniz-Institut für Sozialwissenschaften* is foglalkozni kezdett a félig automatizált indexelő eljárás kérdésével. A folyamat kezdetét a MindServer nevű szoftver megvásárlása jelentette, mely a *sowiport* internetes szakportál számára egy tárgyszavakat javasló rendszert hivatott fejleszteni.

Az intellektuális és az automatikus tartalmi feltárás alapjai

Míg az intellektuális tartalmi feltárás során a tartalom elemzésével és a szaktudás segítségével állapítják meg az adott dokumentum számára legmegfelelőbbnek tartott tárgyszavakat, addig az automatikusan generált tárgyszavak kizárólag a vizsgált szöveg nyelvi formáját veszik figyelembe. Az automatikus eljárások általában olyan visszakeresési modellekben szerepelnek, amelyek a keresőkérdésre egy minden dokumentumot tartalmazó, ám relevancia szerint csökkenő találati halmazt készítenek, tehát az első találat áll a legközelebb a keresett kifejezéshez (Best-Match-Retrieval). Ezzel szemben az intellektuális módon készült indexeket általában olyan rendszerek használják, amelyben csak azokat a dokumentumokat kapjuk találatul, amelyek pontosan tartalmazzák a keresőkifejezést (Exact-Match-Retrieval).

Az automatikus tartalmi feltárás módszerei

Statisztikai módszer

A statisztikai módszer kiinduló gondolata az, hogy minél gyakrabban fordul elő egy adott kifejezés egy szövegben, annál jellemzőbb annak tartalmára. Mivel nem minden kifejezés alkalmas indexelésre, ezért ezeket szelektálni, súlyozni kell.

Nyelvészeti módszerek

A számítógépes nyelvészeten alapuló rendszerek az indexelő kifejezést a nyelv szabályai, morfológia és szintaktikai elemzések alapján választják ki. Kétféle eljárás különböztethető meg: a nyelvi szabályokra épülő, mely a nyelv szabályszerűségeit algoritmusok formájában fogalmazza meg, illetve a szótárra alapuló, mely szerint a nyelvészeti elemzés egy előre meghatározott szótár alapján történik. Mindkét eljárás sok karbantartást igényel.

Fogalomorientált eljárások

A fogalomorientált módszer a szavak jelentéséig, szemantikai szintjéig is eljut. Egy előre meghatáro-

zott szójegyzékkel összehasonlítva választja ki a szöveg jelentést hordozó szavait, melyekből a tárgyszavakat megalkotja. A többértelmű szavak egyértelműsítésére összekapcsolására a statisztikai és a számítógépes nyelvészet módszereit alkalmazza. Ehhez az eljáráshoz is sok háttérmunka szükséges. A bemutatott módszereket általában egymással keverve szokták használni.

Az automatikus tartalmi feltárás értékelése

Az automatikus tartalmi feltárás eredményét visszakeresési teszttel, általában két mutató mentén értékelik ki. A *recall* azt adja meg, mennyire volt teljes körű a visszakeresés, azaz a találatként kapott releváns dokumentumok és a dokumentumgyűjteményben összességében található releváns dokumentumok közötti összefüggést. A *precision* érték a találatok pontosságát adja meg: a találatként kapott releváns dokumentumok, valamint az összes dokumentum viszonyát.

Az automatikus tartalmi feltárás tesztje a SOLIS adatbázisban

A cikk egy vizsgálatot mutat be, mely Krause[1] ún. rétegmodell (Schalenmodell) koncepciójából indul ki, miszerint a *SOLIS (Sozialwissenschaftliches Literaturinformationssystem)* társadalomtudományi szakirodalmi tájékoztató rendszerben a tartalmi feltárás különböző szintjei, rétegei képzelhetők el, a legrelevánsabb szakirodalmat tartalmazó magtól kiindulva. A külső rétegben található kis relevanciájú adatok tartalmát a modell szerint automatizáltan érdemes feltárni. A cikk írója két tesztsorozatot futtatott le, melyek során az adatbázis adatait automatikusan indexelte, majd a kapott eredményeket összehasonlította az intellektuális módon megállapított tárgyszavakkal.

Az 1980-ban elindított SOLIS adatbázis 1945-ig visszamenőleg jegyzi a német nyelvterületen megjelent társadalomtudományi szakirodalom bibliográfiai adatait. A tartalomra egy rövid, esetenként a szerző által készített referátum, tárgyszavak, valamint a társadalomtudományi osztályozás utal. A tárgyszavak a társadalomtudományi tezauruszból származnak, amely 2012-ben 8000 deszkriptort és 5000 nem deszkriptort tartalmazott.

A vizsgálat körülményei

A teszteket a *Recommind* cég által készített MindServer szoftverrel végezték, melyet automatikus tárgyszavazásra és osztályozásra fejlesztettek ki. A szoftver elsősorban a statisztikai módszert alkalmazza, amennyiben egy gyakorló korpusz alapján tanulja meg, hogy milyen valószínűséggel kap egy dokumentum bizonyos tárgyszót és jelzetet. Egyedül a szótövek megállapításánál alkalmazza a nyelvészeti módszert. A szoftver a tárgyszavakat csak a tezaurusz, illetve a társadalomtudományi osztályozás szótárából választhatja ki, így a fogalomorientált megközelítést is alkalmazza.

A szoftver alapját egy algoritmus képzi, amely felismeri a visszatérő koncepciókat és témákat. A szoftver elmenti azt az információt, hogy egy kifejezés előfordult egy bizonyos dokumentumban, és felhasználja a további indexeléshez. Ez kiegészül azzal a valószínűséggel, ami megadja, hogy egy bizonyos kifejezés vagy dokumentum egy bizonyos témához vagy koncepcióhoz tartozik-e. Ezt a hozzátartozást a program másik fontos összetevője, a Support Vector Machine egy vektortérben értelmezi.

A tesztekhez egy 280 dokumentumból álló korpuszt használtak.

Az első tesztsorozathoz az indexelő szoftvert az egész SOLIS adatbankon (kb. 360 000 tétel) „tanították be". Itt nem volt megszabva, hogy a szoftver hány tárgyszót adhat egy dokumentumnak. Mivel így a *recall* érték a *precision*-hoz képest jóval magasabb lett, a pontosság javításának érdekében az adható deszkriptorok számát 10 és 15 között határozták meg.

Míg az első tesztsorozat az egész adatbázisra nézve általános érvényű volt, a második folyamán a rétegmodellnek megfelelően az adatbázis központi és szélső területeit hasonlította össze. A pontosabb eredmények érdekében itt a teszthez kiválasztott tudományterületeken külön-külön tanították be a szoftvert. Így sokkal jobb *precision*-értékek születtek.

Eredmények

A tesztek kiértékeléseként megállapítható, hogy az automatikus indexelés sokkal jobb eredményeket hozott az adatbázis magjához tartozó területeken, mint a szélső rétegekben. Ez azonban adódhat a szélső rétegekhez tartozó tudományok szerkezetéből, illetve abból, hogy a kérdéses tudományterületeken sokkal kevesebb gyakorló dokumentum állt rendelkezésre a szoftver betanításához.

Az adható deszkriptorok számának korlátozása a visszakeresés pontosságát jelentősen megnövelte. Ez a tárgyszavazó munkáját is meggyorsítja, hiszen behatárolt mennyiségű javaslatot kell csak áttekintenie. Így mindenképpen javasolható egy határérték (cut-off level) beállítása.

Bár a szoftver szakterületenkénti betanításával lényegesen jobb eredmények születtek, a pontosság nem lett jobb annál, mint mikor az adható deszkriptorok számát maximalizálták. Ezen kívül ebben az esetben az előkészítő folyamat jóval több energiát emésztett föl, mint amennyivel jobbak lettek a találatok, így a mindennapi gyakorlatban a lefuttatott tesztek alapján nehezen lenne alkalmazható.

Hivatkozás

[1] KRAUSE, Jürgen: Informationserschließung und –bereitstellung zwischen Deregulation, Kommerzialisierung und weltweiter Vernetzung – Schalenmodell. IZ-Arbeitsbericht Nr. 6, Bonn: Informationszentrum Sozialwissenschaften. p. 24.

/KEMPF, Andreas Oskar: Automatische Inhaltserschließung in der Fachinformation. Eine Evaluation zur maschinellen Indexierung sozialwissenschaftlicher Forschungsliteratur. = Information, Wissenschaft & Praxis, 64. köt. 2−3. sz. 2013. p. 96−106./

(Némethné Szivi Zsófia)