Lengyelné Molnár Tünde
Eszterházy Károly Főiskola, Informatika Tanszék
mtunde@ektf.hu

HATÉKONYSÁGNÖVELÉS A KÖNYVTÁRAKBAN SZÓSTATISZTIKAI ELJÁRÁSOK ALKALMAZÁSÁVAL

Bevezetés

Az informatikuskönyvtáros-oktatás fontos területe a tájékoztatás, amelynek alapvető feltétele a különböző szakterületek irodalmának ismerete. Ez viszont nem könnyen megoldható feladat, hisz bármely szakterületet is vizsgáljuk, csupán a folyóiratokban megjelenő legújabb kutatási eredmények publikálásának száma oly magas, hogy lehetetlen mindegyiket áttanulmányozni. Ennek megoldásához a referátumok szerepe fokozatosan felértékelődik, mivel az olvasó, vagy a könyvtáros a cikk teljes áttekintése helyett tizedannyi idő alatt hozzájut a cikk (elvileg) legfontosabb tartalmi anyagához, ezenfelül a referátum áttekintése a redundáns elemek feltárásában is segítséget nyújt. A megnövekedett számú publikációk a dokumentációkészítőket is nehéz feladat elé állítják.

A következőkben szeretném feltárni a referátumkészítés automatizálásának lehetőségeit és korlátait, továbbá áttekinteni azokat az eljárásokat és technikai kivitelezéseket, melyekkel megvalósítható az automatikus referátumkészítés magyar nyelven.

Napjainkban az egyik legfontosabb érték, amely mind a munkánk során, mind a hétköznapokban központi helyre került, a jól informáltság. Ennek eléréséhez a könyvtárak jelentős szerepet nyújtanak. A tájékoztató könyvtárosoktól a felhasználók elvárják a különböző tudományterületek irodalmának ismeretét. Ezzel szemben szinte lehetetlen lépést tartani az egyes szakterületeken megjelenő fejlesztésekkel. Hisz bármely tudományágat is vizsgáljuk, csupán a folyóiratokban megjelenő legújabb kutatási eredmények publikálásának száma oly magas, hogy lehetetlen mindegyiket áttanulmányozni.

A cikkek számát folyamatosan növelik a társadalmi elvárások is, hiszen egy egyetemi, főiskolai oktató vagy bármely szakterülethez tartozó kutató minősítésében nagymértékben figyelembe veszik a publikációk és előadások számát. Ennek következtében a publikációk száma emelkedik, viszont a bennük lévő újdonságtartalom csökken, és egyre nagyobb a cikkekben megjelenő redundancia.

Mindezeket figyelembe véve a referátumok szerepe fokozatosan felértékelődik, mivel az olvasó a cikk teljes áttekintése helyett tizedannyi idő alatt hozzájut a cikk (elvileg) legfontosabb tartalmi anyagához, ezenfelül a referátum áttekintése a redundáns elemek feltárásában is segítséget nyújt.

Úgy gondolom, a referátumok fontosságát senki sem kérdőjelezi meg. Az olvasó számára még az is megvalósíthatatlan, hogy az összes referáló folyóiratot, és annak

minden egyes cikkét figyelemmel kísérje, de legalább lehetőség nyílik arra, hogy nagyobb tájékozottságra tegyünk szert a szakterületünkön. Ne felejtsük, a referátumok áttekintése a legjobb módszer annak eldöntésére, szükséges-e időnket a teljes cikk elolvasásával töltenünk.

Eddig a felhasználó szemszögéből vizsgáltuk a referátum fontosságát, és eljutottunk oda, hogy még ezek áttekintése is gondot okoz a legtöbb szakterületen. Ha az érem másik oldalát vizsgáljuk, a referátumkészítést, ott sem problémamentes kép tárul elénk. A megnövekedett számú publikációk a dokumentációkészítőket is nehéz feladat elé állítják. Hagyományos eszközökkel lehetetlen a teljességet megvalósítani, sőt még megközelíteni is. Az egyetlen megoldásnak az látszik, ha minél nagyobb mértékben bevonjuk a számítógépet a dokumentalista munkájába, és próbálunk minél tökéletesebb eljárásokat kidolgozni, hogy a számítógép önállóan is képes legyen egy cikk, vagy esetleg egy könyv leglényegesebb elemeinek visszaadására.

A továbbiakban a referátumkészítés automatizálásának néhány lehetőségét és korlátját, valamint azokat az eljárásokat és technikai kivitelezéseket tekintjük át, melyek használatával a számítógép automatikusan képes referátumot előállítani.

Dokumentációs válság?

Az említett problémák és megoldások keresése nem az elmúlt 1-2 évben kezdődött. A „dokumentációs válság" kifejezéssel Magyarországon már Szalai Sándor 1963-ban megjelent könyvében is találkozhatunk, mely a gépi kivonatkészítés akkori helyzetét mutatja be. A könyvben találunk egy-két érdekes statisztikai adatot is, mely rávilágít arra, hogy a dokumentumok számának növekedése már az 1800-as években megkezdődött: „1750-ben 12, 1800-ban több mint 90, 1850-ben több mint 900, 1900-ban kb. 9000, 1950-ben kb. 80 000 természettudományos periodika (folyóirat és egyéb időszaki kiadvány) jelent meg a világon". (Szalai, 1963. p. 5.) Ha a friss adatokat szemléljük, 2000-ben a világon kiadott folyóiratok száma 160 000 körül van.[1] A kiadások száma hatványozottan emelkedik.

Ha a „dokumentációs válság" kifejezés már 1963-ban aktuális volt, akkor – a számadatokat szemlélve – napjainkban még inkább az.

Kivonatolás

A referátum kifejezést több – a dokumentum tartalmának visszaadását célzó – feldolgozó eljárás gyűjtőfogalmaként használják. A számítógépes feldolgozás szempontjából a kivonatolás a legjobban automatizálható eljárás. Ezért vizsgáljuk meg ezt a fogalmat egy kicsit részletesebben.

A kivonat bármilyen információs anyag (legyen az írásbeli, vagy szóbeli közlemény) rövidített formában történő visszaadása. Ez a kivonatolás történhet úgy, hogy a közlemény lényegét a kivonatoló saját szavaival írja le. Ez esetben homotopikus (tárgyazonos) közlésről beszélünk. A másik lehetősége a kivonatolásnak, ha a közleményben elhangzott dolgokat, azaz a közlemény tárgyát felsoroljuk. Ezt nevezzük

[1] http://www.sims.berkeley.edu/research/projects/how-much-info/print.html#origflowworld

indikatív (tárgyra utaló) közlésnek. Az indikatív közlés előállítása automatizálható, hisz nincs szükség olyan szakemberre, aki az elhangzott, vagy elolvasott anyagot értelmezi és kiemeli annak lényeges elemeit mindenki számára érthető megfogalmazásban, „csupán" a legnagyobb hangsúlyt kapott elemeket megismétli, kivonatolja. (A két típus nagyon gyakran kombinált formában jelenik meg.)

A fenti osztályozás a kivonat tartalmi megközelítése szerint történt. Ha a nyelv és a logika oldaláról közelítjük meg a kivonat fogalmát, akkor a fent leírt módszereket összefoglaló (summa) és kiválasztó kivonatnak (excerptum) nevezzük. Az összefoglaló kivonat esetén a kivonat készítője a számára fontosnak, hasznosnak tűnő részeket saját megfogalmazásában ismerteti, míg a kiválasztó kivonat esetén a közlemény szövegrészei, vagy annak egységei változatlan formában történő leírásából áll össze a kivonat anyaga. (Szalai, 1963. p. 9–15.)

Statisztikai módszerek

Több statisztikai módszer létezik, melyek között vannak olyanok, amelyeket csak speciális célú elemzések során használunk, és vannak olyanok, amelyek a kivonatolás során elhagyhatatlanok. Ilyen a gyakoriságvizsgálat. A kivonatkészítés automatizálásának első lépése, hogy a benne lévő szavakat önálló egységnek tekintve összeszámoljuk előfordulásaikat. Majd a gyakoriságok szerint rendezzük a kapott adathalmaz, és ez alapján megkapjuk a szöveg statisztikai szótükrét.

Gyakoriság vizsgálatok

Zipf volt az első, aki a szöveg szavainak és szerkezeteinek eloszlásában szabályszerűséget fedezett fel. A vizsgálatokat Joyce Ulysses című regényén végezte és kimutatta, „a regény szavait előfordulási számuk szerint rendezve a kommulatív előfordulásszámok és a bennfoglaló gyakoriságértékek szorzata állandó." (Horváth Tibor–Papp István, 1999. 107. p.)

Ahhoz, hogy gyakoriságvizsgálatokat végezhessünk, a szövegben előforduló szavaknak meg kell keresni a szótövét – ezt típusnak nevezzük –, és ezen szavak különböző megjelenési formáit, előfordulásait – amit jelnek hívunk – fogjuk összesíteni. Az előfordulást gyakoriságuk sorrendjébe rendezzük.

A szótőkeresés elég hosszadalmas és fárasztó munka, ezért ez az a fázis, ahol igyekezni kell a számítógépet bevonni a munkába. Viszont a magyar nyelv esetén ez a legnehezebben megoldható feladat. A számítógépes nyelvészet jelentheti az egyetlen megoldást. Magyarországon a számítógépes nyelvészet fejlődése 1960-ban kezdődött a gépi fordítás korszakával. Ezt az időszakot az orosz–magyar gépi fordítási algoritmus alapjainak kidolgozása jellemezte. A második korszakot (1967–1971) a dokumentációs nyelvészeti csoport munkája alkotja, melynek során kidolgoztak egy saját fejlesztésű, szintaktikai elemző eljárást. A harmadik lexikológiai korszak (1972–1978) eredményei az irodalmár-filológus kutatók igényeinek kielégítésére jöttek létre. Ebben az időszakban kezdődött el a nyelvoktatásban használatható szoftverek fejlesztése, illetve a kvantitatív elemzéseken alapuló gyakorisági szótárak létrehozása a magyar köz- és irodalmi nyelv területén. Ezek az eredmények viszont

egyes személyekhez kapcsolódtak, ugyanis 1972-ben a fővárosban működő Dokumentációs Csoport felszámolásával megszűnt a magyarországi nyelvészeti munka. Az 1979-es újraindulással elkezdődik a negyedik korszak, mely próbálja behozni a 70-es években kiesett tapasztalatok hiányát. Ez időszaktól kezdve Európa-szerte fellendülés tapasztalható a nyelvfeldolgozó rendszerek területén, melynek hatására Magyarországon is elkezdték fejleszteni az MI-nyelvet, és létrejött egy magyar morfológiai elemző alkalmazás is. A 90-es években óriási fejlődés indult meg a személyi számítógépes szoftverek megjelenésével. Az előrelépés egyik állomása a magyar nyelv sajátosságainak megfelelő helyesírás-ellenőrző megjelenése, melynek során algoritmussal írták le szavak összetételét, tehát a szótő és a toldalékok kapcsolódását. A készítő Morphologic cég napjainkra a magyar számítógépes nyelvészet egyik legmeghatározóbb alakjává vált, amikor a Microsoft megvásárolta programjukat. Munkájuk újabb eredményei már a szövegkörnyezetet is vizsgálja, mely kiszűri az irrelevánsnak tűnő értelmezéseket. (Prószéki Gábor, 1989. p. 489–492.) Napjainkra egyre több magyarországi intézmény válik világszerte ismertté számítógépes nyelvészeti munkájával. Az MTA Szegedi Egyetem Mesterséges Intelligencia Kutatólaboratóriumában készült ILP, azaz az Inductive Logic Programming az egész világban kísérleti nyelvészeti alkalmazások egész sorát vonultatta fel.

A fenti eredmények lehetővé teszik a szavak szótövének megkeresését a magyar nyelv esetén is. Viszont ezeket az eredményeket eddig nem alkalmazták a könyvtárinformatika területén.

A szótő megállapításának problémája után a gyakorisági vizsgálatok elvégzéséhez a szavak megszámlálása szükséges, mely egyszerű programozási utasításokkal megoldható.

A gyakorisági vizsgálatok, illetve a kivonatolás elvégzéséhez meg kell határozni a szignifikáns kifejezéseket.

Zipf törvénye szerint a szignifikáns kifejezések a gyakorisági lista adott tartományát jelentik, ami szakterületenként változik, de minden egyes tudományágban igaz, hogy nem a lista eleje és nem is a vége. A szignifikáns szavak listáját megkapjuk, ha a gyakoriság eloszlási függvényére rávetítjük a tudományterületre jellemző tapasztalati úton meghatározott Gauss-görbét. (Horváth Tibor–Papp István, 1999. p. 56.).

Magyar szövegeket tekintve kevés tudományágnak létezik gyakorisági szótára, mely alapján a Gauss-görbe felállítható lenne. Jelenleg a Magyar Tudományos Akadémia foglalkozik szógyakorisági szótárak összeállításával.

Ha számítógéppel szeretnénk meghatározni a releváns helyekhez vezető kifejezéseket, akkor mindenképpen figyelembe kell venni azt, hogy vannak-e a szövegben olyan szópárok, illetve szóhármasok melyek többször fordulnak elő. Ez az elgondolás Luhn-tól ered, aki 1951-ben jelentette meg elképzelését. A szomszédos szavakat, illetve szóhármasokat, a triviális szavak elhagyása után kell vizsgálni, majd egy súlyozás bevezetésével jutunk el a releváns szövegrészekhez. Ennek módja, hogy a két- vagy többtagú nem triviális szóelőfordulások magasabb súlyt kapnak, mint azok egyszeres előfordulásai. A súlyok megalkotása után dönteni kell arról, hogy milyen egységeket akarunk visszakapni releváns helyként: mondatot vagy bekezdést. Ezután történik az automatizálás: hozzárendelünk egy számértéket a választott egység-

hez a súlyok alapján, és a legmagasabb számértékkel rendelkező mondatokat, vagy bekezdéseket adjuk vissza eredményként.

Nehézségek a gépi kivonatkészítés során

- Gondot okoz a szópárok, szóhármasok keresése, hiszen az eredeti szövegben nem feltétlenül lesznek egymás mellett a releváns párok, mert triviális szavak elválasztják őket egymástól. A lépés automatizálása megoldható, hiszen már ma is létezik több olyan szoftver, amely lehetővé teszi, hogy a számítógép képes legyen az egymástól néhány szónyi távolságra lévő kifejezések keresésére.
- Sok mondatban az alany megjelölése az előző mondatokban kiírt személy(ek)re, esemény(ek)re történő utalás formájában jelenik meg. Ezek kezeléséhez a szövegtan eredményeit kell felhasználnunk.
- A tökéletesebb eredmények eléréséhez szükséges a statisztikai vizsgálatokat kiterjeszteni, és figyelembe venni, a szignifikáns szavak első előfordulását, és ezt súlyozással jutalmazni.
- További fejlesztések közé tartozna, ha nemcsak a szavakat vizsgálnánk, hanem figyelembe vennénk a mondat elhelyezkedését is a bekezdésen belül. Ugyanis a szerzők a bekezdés első mondataiban általában megjelölik mondanivalójuk tárgyát, záró mondatában pedig adnak egy összegzést. Ezek hasznos mondatok, így megéri a súlyozásnál ezt is figyelembe venni. A gondolatmenetet folytatva a bekezdés helyét is hasonló okokból vizsgálhatnánk.
- Nagymértékben csökkenti a hatékonyságot, ha a szerző igyekszik változatos kifejezéseket használni, és ugyanazt a dolgot, vagy személyt különböző elnevezésekkel illeti.

Összegzés

Befejezésül hangsúlyoznám, hogy ez az eljárás csak diszkurzív szövegek esetén használható, amikor is a szöveg egy témakört tárgyal, és azt tényközlő megállapításokkal teszi, nem pedig a legváltozatosabb irodalmi stílust használva, továbbá a szerző következetes a szóhasználatban és mondanivalójának tagolásában. Általánosságban elmondhatjuk, hogy a Luhn-féle módszer hatékonyabban alkalmazható a tudományos közlések, jelentések esetén, mint egy választékos irodalmi stílussal megírt mű esetén.

Magyarországon eddig nem jelent meg automatikus referátumkészítő program, eddig csupán az igény tapasztalható.

Irodalomjegyzék

Antal László: A tartalomelemzés alapjai. Budapest, 1975, Tömegkommunikációs Kutatóközpont.
Horváth Tibor–Papp István: Könyvtárosok kézikönyve 1. Budapest, 1999, Osiris. /Osiris kézikönyvek/
Horváth Tibor–Papp István: Könyvtárosok kézikönyve 2. Budapest, 2001, Osiris. /Osiris kézikönyvek/
Murray R. Spiegel: Statisztika. Budapest, 1995, Panem–McGraw–Hill /Schaum-könyvek/
Pietil, Veikko: Tartalomelemzés. Budapest, 1979, Tömegkommunikációs Központ
Prószéky Gábor: Számítógépes nyelvészet. Budapest, 1989, Számítástechnika-alkalmazási Vállalat.