Kódrendszer dokumentumok egyértelmű azonosítására

Az ADONIS program során merült fel a feladat, hogy egyértelmű kapcsolatot kell teremteni a dokumentummásolatokra vonatkozó igények és az ADONIS program CD-ROM lemezein tárolt dokumentumok között. Ha a másolatra vonatkozó igény bibliográfiai adatbázisra alapozva gépi úton érkezik, akkor a kapcsolatteremtésnek automatikusnak kell lennie, de a többi igényről is gépi úton kell eldönteni, kielégíthetők-e az ADONIS lemezekről, és ha igen, akkor meg kell találni a dokumentumok ADONIS azonosítóját, végül az igényt automatikusan továbbítani kell az ADONIS munkaállomásra. Ennek a feladatnak a megoldására indult a DOCMATCH program.

A DOCMATCH keretében kidolgozott azonosítási rendszer a bibliográfiai adatbázisokra alapozott, gépi úton érkező igények jó minőségű bibliográfiai leírására készült. Így külön meg kellett vizsgálni, hogy hogyan válik be manuális adatbevitel esetén.

Az azonosítás alapja a DOCMATCH keretében kidolgozott rendszerben az *USBC* (Universal Standard Bibliographic Code = univerzális szabványos bibliográfiai kód) névre keresztelt kódrendszer. Az ADONIS lemezeken tárolt minden dokumentumra tartalmazza ezt a kódot az adatbázis indexe, és minden beérkező másolatigényre számítógép építi fel ezt a kódot.

Az USBC 16 karakterből áll, és részei a következők:

1 karakter a megjelenés évéből,
6 karakter az első szerző vezetéknevéből,
2 karakter a paginációból,
7 karakter a címből.

Az évszám karaktere az ábécé annyiadik betűje, amennyi az évszám 26-tal történő osztásának maradéka.

Az első szerző vezetéknevéből képezett hat karakter a névben *legritkábban* előforduló betűkből áll, a kis- és nagybetűk megkülönböztetése nélkül. Az azonos gyakorisággal előforduló betűk ábécésorrendben állnak. Ha nincs ki a hat betű, a hiányzó helyeken csillag áll. (Pl. a Balassa név kódja: BLSA**, a Yannakoudakis névé DIKSUY. – A ref.) A vezetéknév egyértelmű kiválasztására szabályrendszer szolgál. Lista készült például a vezetéknév előtt elhagyandó előtétekről: Mc, Mac, Von, Van, D', De, Den stb. Ha az előtétek elhagyásával elfogyna a név, egyet vissza kell lépni. Ha nem választható ki egyértelműen a vezetéknév (pl. kínai nevek), akkor a teljes nevet használják. Világos, hogy a nehezen kezelhető nevek kódolása egyeztetési nehézségekre vezet, de az a vélemény alakult ki, hogy ez a megoldás még mindig sokkal kevesebb problémát okoz, mint bármelyik más*.

A pagináció két karaktere a kezdő oldalszám két utolsó jegye, szükség esetén csillaggal egészítve ki két karakterre.

A cím hét karaktere a szerző kódjához hasonlóan a címben legritkábban előforduló alfanumerikus karakterekből áll, azonos gyakoriság esetén 0–9, A–Z sorrendben.

A kód egyes részeinek a sorrendje megfelel annak, hogy a bibliográfiai leírás pontossága a papíron érkező másolatigényekben romlik. A kérés és a tárolt dokumentum kódjának egyeztetésekor teljesnek tekintendő az egyezés, ha a kód minden eleme egyezik, jónak, ha a cím kódrésze kivételével megvan az egyezés, és gyengének, ha csak az évszám és a szerző kódja egyezik.

A rendszer tesztelése során az online szolgáltatóközpontokon (pl. DIALOG) keresztül érkező igényekre majdnem százszázalékos volt az egyeztetés sikere. A másolatszolgáltatás teljes automatizálása azonban itt is nehézségekbe ütközik, például az adatrekordok mezőkre tagolása a keresési eredményekben nem egységes, és a gépi program számára nem mindig világos. Ezért például a DIALOG rendszer mind a 23 számba jövő adatbázisára külön-külön be kellett vinni a programba a kinyomtatott keresési eredmény rekordszerkezetét. Jobb volt az eredmény, ha a megrendelő nem a nyomtatási formátumra, hanem a letöltési formátumra (tagged output) alapozta a kérését. Még ilyenkor is nehézséget okozhattak azonban az ADONIS felépítésekor elkövetett gépelési hibák és a speciális karakterek (pl. görög betűk) nem egységes kezelése. Ha teljes egyezést nem lehetett találni, részlegesen egyező dokumentumok közül emberi döntéssel kellett választani.

További nehézséget okoz az ADONIS-igények kis aránya az összes igény között. Bár az orvosi-biológiai témakört az ADONIS jól lefedi, egy több száz másolatigénylést tartalmazó adatállományban mégis csak 2–3 olyan igény található, amely az ADONIS rendszerből elégíthető ki. Ez az alacsony arány nagyon nagy feldolgozási időre vezet. A gyorsítás érdekében erősen automatizálni kellett a folyamatot, vagyis, ha nem talált a program egyezést, akkor automatikusan a hagyományos másolatszolgáltatáshoz terelte az igényt, ahelyett, hogy az operarátor döntésére bízta volna. Az operátorra csak a sikeres egyezések megerősítése maradt, ez a minimálisra leszorított interaktivitás nagy adatállományok gyors feldolgozását tette lehetővé. A gyorsaság ára viszont az, hogy a csekély hibával beadott, az operátor számára még felismerhető ADONIS-igények is a manuális feldolgozáshoz kerültek.

---

* A szerzők nem látszanak felmérni a *transzliterálás* többértelműsége okozta problémákat. Már a német ä, ö, ü hol a, o, u, hol ae, oe, ue formájú transzliterálása és egyes skandináv betűk többféle kezelése is egyeztetési nehézségeket okozhat. A rendszer cirill betűs folyóiratokra való kiterjesztését azonban végképp meggátolhatja a teljesen rendezetlen transzliterálás. – A ref.

A DOCMATCH program hasznos melléktermeke volt a feldolgozási hibák folytán létrejött duplumrekordok feltárása az ADONIS-állományban, mivel ezek azonos USBC-t kaptak.

A gépi adathordozón érkező másolatigények között is sok volt olyan, amelyik nem adatbázison alapszik. Ezek természetszerűen sokkal pontatlanabbak, mint az adatbázison alapulók, és kevesebb információt tartalmaznak. A címet gyakran csak csonkolva közlik, a folyóirat címét pedig messze nem szabványos formában rövidítve. Ezeknek a problémáknak a hatását a DOCMATCH feldolgozásra nem a beérkező igények állományán vizsgáltuk. Abban ugyanis összeadódnak a nem adatbázison alapuló igények problémái és a fentebb említett rekordformátum-problémák. Helyette a Bradfordi Egyetem Könyvtára (University of Bradford Library) könyvtárközi kölcsönzési rendszerének régi adatállományai szolgáltak a vizsgálat alapjául, amelyek formailag teljesen rendben vannak, így nagy adatállományokon végzett vizsgálatokra adnak módot.

A kísérlet első lépésében teljes egyezést kívántunk meg, és ekkor egyáltalán nem kaptunk találatokat. Amikor azonban a második lépésben megelégedtünk már a jó egyezéssel, vagyis a legrosszabb minőségű rész, a címből kapott kódrész egyezését nem kívántuk meg, meglepően sok találatot kaptunk, mégpedig sok esetben egy igényre egy találatot, néhány esetben kisszámú lehetséges találatot, amelyek közül könnyű volt kiválasztani az igazit. Egyes kérésekre egy vagy több hamis találatot kaptunk csak, ezek azonban szinte ránézésre kizárhatók voltak. Bár az operátor számára könnyű feladat volt a valódi és a hamis találatok elkülönítése, ennek a döntésnek az automatizálása nagyon nehéz feladat. A nehézséget a cím lehetséges idézési hibáinak sokfélesége okozza: csonkolás, rövidítések, átfogalmazások vagy ezek kombinációi. Az operátor például könnyen azonosítja a következő két címet: "The origins of the Second World War" és "WWII, Origins", de az egyszerű, betűről betűre történő gépi összehasonlítás különbözőknek tekinti őket.

Miután világossá vált, hogy a cím nagyon rosszul használható azonosításra, megvizsgáltuk más bibliográfiai elemek lehetséges felhasználását. Az ISSN-t azért kellett elvetnünk, mert az nagyon ritkán szerepel a beérkező igényekben. A folyóiratcím hasonló problémáktól szenved, mint a cikkcím: rövidítés, csonkolás. Ezt az elemet azonban nem kell teljesen elvetnünk az összehasonlításból. A valódi és a hamis találatok közötti döntésben például nagyon hasznos az az egyszerű módszer, hogy összehasonlítjuk a folyóiratcím első betűjét. A kötetszám, folyóiratszám és rész adatai kevésbé használhatók, mint várnánk, mivel a számozási rendszerekben sok az eltérés. Az ADONIS rendszerbe bevitt ilyen adatok például sokszor inkorrektnek bizonyultak a "rendes sorrenden" kívüli folyóiratszámok (pl. supplementumok), a több részre osztott folyóiratszámok és az összevont folyóiratszámok esetén. Az 1–2. szám például időnként mint 12. szám került be. Az sem egységes, hogy mennyit adnak meg ezekből az adatokból a másolatigénylők. A kötetszámot viszont szinte mindig megadják, ezért a hamis találatok kizárására legjobbnak a folyóiratcím első betűjének és a kötetszám utolsó számjegyének az egyeztetése bizonyult. Tovább növelhető az egyeztetés biztonsága, ha a kezdő oldalszámból felhasznált számjegyek számát kettőről háromra növeljük, mert a folyóiratok jelentős része kötetenkénti oldalszámozást alkalmaz, így nagy a háromjegyű kezdő oldalszámok aránya.

Szerettük volna összehasonlítani az USBC használatának hatékonyságát más hasonló kódokéval, az *ISO BIBLID* kódéval és a *NISO SAID* kódéval. A problémák azonban olyan súlyosak voltak, hogy egyszerűen nem tudtuk ezeket a kódokat generálni. Mind a BIBLID, mind a SAID a következő adatstruktúrán alapszik: ISSN, dátum, számozás, paginá-ció. Mint már említettük, az ISSN a nem adatbázison alapuló másolatigényekből általában hiányzik. A SAID által igényelt teljes dátum nincs rajta az ADONIS rekordokon. A számozásban a már említett problémákon ("rendes sorrenden" kívüli, megosztott és összevont számok) kívül egyes folyóiratok szokatlan számozási gyakorlata is gondot okoz. A *The Lancet* például a következő formulát alkalmazza: "Vol. II for 1989". Ha sikerül generálni a kódot, akkor is bajt okoz, hogy mindkét rendszerben azonos lesz a kódja az egyazon folyóiratoldalon kezdődő két cikknek. A kétértelműség azzal hárítható el, ha az amúgy is már túl hosszú kódokat még kiegészítjük a címből képezett résszel, de a BIBLID kódra vonatkozó ISO-szabvány a cím egyeztetésre való felhasználását explicite megtiltja.

/AYRES, F. H.– HUGILL, J. A. W.– RIDLEY, M. J.– YANNAKOUDAKIS, E. J.: DOCMATCH: automated input to ADONIS. = Interlending and Document Supply, 18. köt. 3. sz. 1990. p. 92–97./

(Válas György)

A "szabadpolcos" számítógépek védelme

Egyre több az olyan számítógép, amelyhez szélesebb közönség férhet hozzá, például egy könyvtár olvasói. Ilyen gépek szolgálhatnak a CD-ROM adatbázisokban vagy más mikroszámítógépes adatbázisokban történő keresésre, hipertext, hipermédia rendszerek használatára, szakértő rendszerek futtatására