002.513.5

A DOKUMENTÁCIÓS RENDSZEREK KISÉRLETI ÖSSZEHASONLÍTÁSÁNAK MÓDSZEREI

Vásárhelyi Pál

A dokumentációval foglalkozó szakembereket világszerte foglalkoztatja az az alapvető fontosságu kérdés, hogy a számos dokumentációs rendszer közül melyik a legjobb, illetve pontosabban, hogy adott helyzetben, adott feladat megoldására melyik rendszert célszerü alkalmazni. Elvi síkon számosan foglalkoztak a dokumentációs rendszerek és azok fő tényezőinek összehasonlitásával, de nem végeztek gyakorlati kisérleteket és nem vonták be a legilletékesebbeket, magukat a felhasználókat az eredmény elbirálásába. Uttörő munkát ezen a téren CLEVERDON kutatócsoportja és a clevelandi /USA/ Western Reserve University /továbbiakban: WRU/ munkatársai végeznek A.GOLDWYN és A.REES vezetésével. CLEVERDON az ugynevezett Aslib Cranfield kutatási program keretében folytatja vizsgálatait, mig a WRU-n rendszerösszehasonlító laboratoriumot állitottak fel, hogy a szakirodalom feldolgozásának és az információ visszakeresésének különböző módszereit sorra kipróbálják, összehasonlitsák és megállapitsák, hogy a módszerek végeredményben hogyan elégitik ki magának a felhasználónak az igényeit.

Ennek érdekében azt a célt tüzték maguk elé, hogy:

- meghatározzák az információ-visszakereső rendszer lényeges összetevőit,
- megállapitsák, melyek azok a tényezők, amelyek elsősorban befolyásolják a rendszer teljesitőképességét,
- kidolgozzanak oly módszereket, melyek lehetővé teszik a rendszer teljesitőképességének mérését, és
- ellenőrizzék a kisérletek utján kapott eredményeket.

A kiinduló feltevés szerint, melyet kisérletileg bizonyitani kivánnak, a rendszer teljesitőképességét elsősorban az alábbi tényezők befolyásolják:

1. A beszerzés. A beszerzés során követett politika, a beszerzett dokumentumok fajtái, minősége stb. meghatározzák a do-

kumentumgyüjtemény tartalmát és igy döntően befolyásolják a rendszer teljesitőképességét.

2. Az input forrása. Az információk visszakeresése szempontjából alapvető fontosságu, hogy a dokumentációs iroda miből indul ki a további feldolgozás során: cimeket, referátumokat, vagy teljes szövegeket választ inputként az indexeléshez.

3. Az indexelés módja. Az indexelésre felhasznált fogalmak és szabályok összessége erősen befolyásolja egy rendszer működését.

4. A kódolás. Az információvisszakeresés szempontjából nagy jelentőséggel bir, hogy az indexelés során felhasznált fogalmakat milyen szimbolikus formában ábrázolják.

5. A dokumentum gyüjtemény szervezésének módja. Kihatással van a végeredményre az is, hogy milyen sorrendben, milyen rendszerezési elv szerint tárolják az információkat.

6. Kérdés-analizis. A nemdokumentációs szakember által feltett kérdést a dokumentációs rendszer nyelvére kell leforditani, az indexelés során felhasznált fogalmakat és szabályokat kell a kérdés végleges megfogalmazásakor is használni.

7. A visszakeresés stratégiája. Az információ visszakeresésére minden rendszerben más konkrét eljárást alkalmaznak, ami ugyancsak érezteti hatását az adott válaszban.

8. A válasz /output/ formája. A feltett kérdésre a választ a megfelelő dokumentumok átadására, referátumok vagy cimek szolgáltatásával lehet megadni.

A rendszerek működésének fenti tényezőit másképpen is csoportosithatjuk és megkülönböztethetünk az információknak későbbi visszakeresés céljából történő megjelölésével, és a megjelölt anyag visszakeresésével kapcsolatos tényezőket. Az első csoportba tartoznak: a kódolás, az input forrása, az indexelés módja, az output formája. A második csoportba tartoznak: beszerzés, kérdés-analizis, a gyüjtemény szervezésének módja és a keresés módja.

A rendszerek továbbá aszerint különböznek egymástól, hogy mely szakterületet akarnak kiszolgálni, a felhasználók mely csoportja részére dolgoznak és hogy mekkora a munkájuk alapját képező dokumentum-gyüjtemény.

A rendszerek értékelésénél abból indulnak ki, hogy egy rendszer teljesitőképessége hatékonyságának és hatásfokának függvénye. A rendszer hatékonyságának mutatószámával azt mérik, hogy a rendszer menynyire képes annak a feladatnak ellátására, amire tervezték, mig a hatásfok mutatószáma ezen feladat megoldásának költség-kihatásait méri.

A laboratoriumi vizsgálatok során egyelőre csupán a hatékony-

ság vizsgálatával foglalkoznak. A hatásfokra vonatkozólag mindössze annyit állapitottak meg, hogy az az időnek és költségtényezőnek függvénye.

Az ideális, maximális hatékonysággal dolgozó információ-visszakereső rendszer jellemzője, hogy mindazokat a dokumentumokat kiemeli a dokumentum-tárból, amelyek a feltett kérdésre helyes, találó választ adnak, de nem emel ki egyetlen olyat sem, amely téves, vagyis a kérdéshez nem kapcsolódó anyagot tartalmaz. Ezzel kapcsolatban azonban egy lényeges szempontra kell felhivni a figyelmet. A szakember, aki a dokumentációs irodához fordul, valamely problémára keresi a megoldást. Ezt a problémát szavakba önti és igy adja a dokumentátor tudtára. A megfogalmazott kérdés azonban nem feltétlenül fedi magát a problémát, annál is inkább, mert a kérdező gyakran maga sem tudja pontosan, hogy mi is a kivánsága. Következésképpen az adott válasz is kétféle lehet. Ha pontos és helyes választ ad a feltett kérdésre, találónak nevezzük, ha pedig magát az eredeti információs igényt elégiti ki, a problémát oldja meg, akkor megfelelőnek mondjuk. A találó válasz nem feltétlenül megfelelő, és forditva, a megfelelő válasz sokszor nem találó. A találó és megfelelő válasz közti kapcsolat vizsgálata nagyfontosságu és többek között az ugynevezett kérdéselemzés módszerének kidolgozásához vezetett. A rendszer hatékonyságának elemzésekor kiindulhatunk mind a találó, mind pedig a megfelelő dokumentumokból, bár az utóbbi meghatározása természetesen nehezebb.

Ezek után nézzük meg, hogyan mérhető és számitható egy információ visszakereső-rendszer hatékonysága. A hatékonyság mérésére az érzékenység és a szelektivitás mutatószámát használják. Az érzékenység annak feltételes valószinüsége, hogy a dokumentumtár egy tagját a rendszer kiemeli abban az esetben, ha az találó /megfelelő/. A szelektivitás ezzel szemben annak feltételes valószinüsége, hogy a dokumentumtár egy tagját a rendszer nem emeli ki abban az esetben, ha az nem találó /nem megfelelő/. A valószinüségszámitás szabályai szerint mind az érzékenység, mind a szelektivitás értéke 0 és 1 között lehet.

A hatékonyságot a fenti két tényező függvényében a következőképpen határozták meg:

$$\text{hatékonyság} = \text{érzékenység} + \text{szelektivitás} - 1$$

A hatékonyság értéke ennek értelmében +1 és -1 között lehet. A maximális +1 értéket abban az esetben veszi fel, ha mind az érzékenység, mind a szelektivitás értéke 1. Ez akkor áll fenn, ha a rendszer kiemeli mindazokat a dokumentumokat, melyek találóak /megfelelőek/ és csak ezeket a dokumentumokat emeli ki. Ez a lehető legjobb eredmény. Másrészt a minimális, -1 értéket a hatékonyság akkor veszi fel, ha mind az érzékenység, mind a szelektivitás 0. Ez abban az esetben következik be, ha a rendszer kiemeli mindazokat a dokumentumokat, melyek nem találóak /nem megfelelőek/, de csak ezeket emeli ki. Ez a lehető legrosszabb eredmény. A hatékonyság 0 lesz abban az esetben, ha annak valószinüsége, hogy egy találó dokumentumot kiemel a rendszer, egyenlő annak valószinüségével, hogy nem találó do-

kumentumot emel ki. Ekkor tehát kívánatos és nem kívánatos anyagot egyenlő valószínüséggel kaphatunk, csakugy mintha véletlen alapján választanánk ki a dokumentumokat a gyüjteményből. A hatékonyság pozitiv, ha a találó dokumentumok kiemelésének valószínüsége nagyobb, mint a nemtaláló dokumentumok kiemelésének valószínüsége, ellenkező esetben negativ.

A hatékonyság értékelésénél mindeddig egyenlő sullyal vettük figyelembe az érzékenységet és a szelektivitást. Reálisabb eredményt kapnánk, ha e két tényezőt különböző együtthatók segitségével sulyoznánk. Ezen együtthatókat azonban minden alkalommal az adott feladattól és helyzettől függő más-más értékben kellene meghatározni, és az általános jellegü vizsgálatot nem zavarja, ha a két együttható értékét 1-nek tekintjük.

CLEVERDON a rendszer hatékonyságát befolyásoló tényezők közül csak az indexelés módjának hatását vizsgálja. Célja az, hogy meghatározza, melyek magának az indexelési folyamatnak fő lépései, és mi azok kihatása az információ visszakeresés hatékonyságára. Az első feladat ezzel kapcsolatban az, hogy biztositsa az egyéb tényezők hatásának kiszürését és laboratoriumi vizsgálatokra alkalmas, pontosan körülhatárolt és ismert kisérleti dokumentumgyüjteményt állitson összesze. A kisérleti dokumentum-gyüjteményt ezért kutatási jelentések alapján hozta létre a következő két alapfeltevésből kiindulva:

Valamely kutatás eredményét közlő jelentés kiindulópontja mindig egy kérdés, egy probléma, amire a kutató választ keresett és talált;

a jelentések irodalomjegyzékében szereplő müvek valami anyagot kell, hogy tartalmazzanak a kutatás kiindulópontjaként szolgáló kérdéssel kapcsolatban.

Ezen feltevések alapján kérdőivet küldtek ki mintegy 400 jelentés szerzőjének, melyben felkérték, hogy a lehető legpontosabban határozza meg, mi volt kiinduló problémája, és melyek voltak a munkája során felmerült további kérdések, melyekre jelentése ugyancsak választ ad. A kérdőiven feltüntették továbbá az illető szerző irodalmi hivatkozásainak jegyzékét és megkérték, hogy ossza be a cikkeket a következő csoportokba:

1. Olyan hivatkozás, mely teljes választ ad egy kérdésre. Nyilvánvaló, hogy ez nem a kiinduló problémára, hanem a felmerült kiegészitő kérdésekre vonatkozik.

2. Olyan hivatkozás, amely igen közel áll a kiindulási kérdéshez, és melynek hiányában a kutatást vagy el sem tudta volna végezni, vagy pedig igen sok többletmunkára lett volna szükség.

3. Olyan hivatkozás, mely jelentős volt, de csak mint általános háttér, vagy a munka bizonyos fázisa szempontjából.

4. Minimális értékü hivatkozás, mely pl. csak a történelmi visszapillantást szolgálta.

5. Értéktelen hivatkozás.

A kérdőiv alapján a kutató által feltüntetett kérdések és a hivatkozásban szereplő dokumentumok képezik a vizsgálatok alapját, a kisérleti kérdés- és dokumentumgyüjteményt. Az eredeti tanulmányokat a gyüjteményből kizárták, mert tulnagy a kérdések és a dokumentumok közötti korreláció. Igy 1500 dokumentumból és 400 kérdésből álló gyüjteményt hoztak létre és biztositották, hogy minden kérdés esetében legyen 1-2 olyan dokumentum, mely arra többé-kevésbé pontos választ ad. Természetesen lehetséges, sőt valószinü, hogy a gyüjteményben szereplő dokumentumok közül olyanok is kapcsolódhatnak egy kérdéshez, melyeket a szerző nem emlitett hivatkozásai között. Ezért a teljes gyüjteményt valamennyi kérdés szempontjából analizálni kell. Ezt a munkát külső szakemberek bevonásával végezték el. Igy végeredményben mélyrehatóan ismert dokumentumgyüjteményt kaptak, melynek minden tagjáról pontosan tudják, hogy egy adott kérdéshez milyen mértékben kapcsolódik. Ha a továbbiakban mindig azonos kódolási és visszakeresési módszert alkalmaznak, a hatékonyság már csupán az indexelés módjától függ, és várható, hogy az indexelés módja és a hatékonyság közötti kapcsolatra vonatkozólag értékes megállapitásokat tudnak tenni.

A WRU rendszer-összehasonlitó laboratoriumában az információ visszakeresés hatékonyságát befolyásoló tényezők közül már négyet vesznek alapos vizsgálat alá. Ezek: az input formája, az indexelés módja, a kódolás és a válasz formája. A többi tényezőket egyelőre figyelmen kivül hagyják, de szigoruan állandó formában tartják azokat, hogy a végeredményt, a vizsgált tényezők változtatása alapján végzett összehasonlitást ne befolyásolják. A kisérletek folyamán a következő lehetőségeket analizálják:

1. Input formája: cim, referátum, teljes szöveg.

2. Indexelés módja: telegrafikus referátum, gépi- és kéziuton meghatározott kulcsszavak, meta-nyelv, tárgyszavak.

3. Kódolás: természetes /angol/ nyelv, szemantikus kód.

4. A válasz formája: cim, referátum, teljes szöveg.

Egy-egy tényező változtatásának hatását az információ visszakeresés hatékonyságára ugy határozzák meg, hogy kiválasztanak teljes dokumentumtárukból egy olyan csoportot, mely statisztikai vizsgálatok céljára már elég nagyszámu dokumentumot tartalmaz, de még elég kicsi ahhoz, hogy a legkülönbözőbb módszerekkel ismételten fel lehessen dolgozni. Ez a dokumentum csoport képviseli tehát az egyszerüség kedvéért a teljes dokumentumtárat. A kiválasztott összes dokumentumot feldolgozzák olymódon, hogy ugyanazt az input-formát, kódolást és indexelési módot alkalmazzák mindegyiknél. Azután a dokumentációs központnak feltett kérdések közül kiválasztanak néhányat, a reprezentativ dokumentumcsoportból információ-visszakeresést vé-

geznek és választ adnak ugyancsak egy meghatározott formában. A kérdésnek megfelelő szakterület legjobb ismerői közül kiválasztanak egy csoportot és azok ellenőrzik az összes /tehát nem csupán a visszakeresett/ dokumentumokat. Az ellenőrzés alapján megállapitják, hogy melyek a reprezentativ dokumentumcsoportból az adott kérdésre találó választ adó dokumentumok és táblázatokat készitenek, melyben az alábbi adatok szerepelnek:

a visszakeresett találó dokumentumok száma

a visszakeresett nem-találó dokumentumok száma

a vissza nem keresett találó dokumentumok száma

a vissza nem keresett nem-találó dokumentumok száma.

A táblázat felhasználásával kiszámitják:

1. az alkalmazott rendszer érzékenységét, oly módon,hogy a visszakeresett találó dokumentumok számát elosztják a visszakeresett összes dokumentumok számával, és

2. a szelektivitást, oly módon, hogy a vissza nem keresett nem találó dokumentumok számát elosztják az összes nem találó dokumentumok számával;

3. a fenti két tényező alapján az alkalmazott rendszer hatékonyságát.

Ezután a vizsgált négy tényező közül hármat változatlanul hagyva, de pl. az indexelési módok közül egy másikat választva ujból elvégzik a teljes kisérletsorozatot meghatározva végül ebben az esetben is a hatékonyságot. Mivel a kisérletek során igy mindig csak egyetlen tényező változik, a hatékonyság számértékében mutatkozó eltérések elég jól mutatják az illető tényező befolyását.

A WRU kisérletsorozatát 600 dokumentummal végzi, melyeket véletlen alapján válogattak ki a ragályos betegségekkel foglalkozó gyüjteményből. A kiválasztott dokumentumokat eddig három különböző indexelési módszerrel dolgozták fel:

a/ telegrafikus referátumokat készitettek róluk, s ezeket mágnesszalagra rögzitették,

b/ a cimek alapján elektronikus számitógéppel kulcsszavas /KWIC/ indexet készitettek, és

c/ tárgyszavas indexet hoztak létre a szakterület referáló-folyóirata tárgyszavai alapján.

A feldolgozást az emberi tényező befolyásának lehető kis értékre történő leszoritása céljából három olyan dokumentátor végezte, akik azonos koruak, azonos képzettségüek, sőt arra is ügyeltek,hogy társadalmi hátterük is hasonló legyen. Munkájuk ritmusát ugyancsak azonos értéken tartották. - Eddig tehát az indexelés módját változtatták. A többi tényezőt változatlanul hagyva információ visszakere-

sést végeztek az orvosi dokumentációs központnak feltett kérdések közül véletlen alapján kiválasztott néhány kérdéssel kapcsolatban.A dokumentumokat azután szakemberekkel ellenőriztették, akik megállapitották, hogy azok találóak, közelállóak, vagy egyáltalán nem találóak a kérdés szempontjából. A hatékonysági vizsgálatok első számszerü eredményei a közeljövőben várhatók.

A későbbiek folyamán hasonlóképpen fogják vizsgálni, hogy:

a/ hogyan befolyásolja a végeredményt, ha az indexelést a teljes szöveg, vagy csupán referátum figyelembevételével végzik el;

b/ hogyan dolgozza fel ugyanazt a dokumentumot más-más dokumentátor?Hány tárgyszóval jellemzi és melyek azok a tárgyszavak, amelyeket legjellemzőbbeknek tartanak a különböző dokumentátorok;

c/ hogyan dolgozza fel ugyanaz a dokumentátor ugyanazt a dokumentumot különböző időpontokban /néhány hónapos időközöket vizsgálva/,

d/ mi a kihatása, ha a felhasználóknak a kérdésre csupán cimet, vagy referátumot, illetőleg, ha teljes szöveget adnak, stb.

Előreláthatólag még évekig el fog tartani, mig csupán az előbbiekben emlitett négy fő tényezővel és azok részleteivel az információ-visszakeresés hatékonyságára gyakorolt befolyását megállapitják.

Érdemes felfigyelni arra, hogy a WRU rendszerösszehasonlitó laboratoriumának munkája a legszorosabb kapcsolatban van az egyetemen folyó oktatással. A hallgatók sajátmaguk dolgoznak fel egy kis dokumentum-csoportot a legkülönbözőbb indexelési és kódolási módszerek felhasználásával, kézi lyukkártyák, valamint az egyetem elektronikus számitógépei igénybevételével információ-visszakeresést végeznek konkrét kérdésekkel kapcsolatban és kiértékelik a kapott eredményeket. Ezzel egyrészt megfelelő gyakorlatra tesznek szert a dokumentáció különböző eszközeinek használatában, másrészt a laboratorium kutatásait is előbbreviszik, hiszen vizsgálataik gyakran érdekes összefüggésekre, jelenségekre hivják fel oktatóik figyelmét.

*

*** ***

IRODALOM

CLEVERDON, C.W. - MILLS, J.: The Testing of Index Language Devices. Aslib Proceedings, XV. 1963.ápr. p.106-130.

REES, A.M.: Sematic Factors, Role Indicators et alia, Eight Years of Information Retrieval at Western Reserve University, Aslib Proceedings, XV. 1963.dec. p.350-363.

GOLDWYN, A.J.: The Place of Indexing in the Design of Information Systems Tests. Automation and Scientific Communication. II.rész. Washingon, American Documentation Institute, 1963. p.321-322.

GOFFMAN, W. - NEWILL, V.A.: Comparative Systems Laboratory, Technical Report No.2. Cleveland, Western Reserve University 1964.jul.

GOFFMAN, W.: Final Report on Theory of Documentation and Search Strategy, Cleveland. Western Reserve University, 1963.márc.

GOLDWYN, A.J.: The comparative systems laboratory in the health sciences at Western Reserve University Congresso Rassegna Internazionale Documentazione Scientifico Tecnica. Roma, 1964. február.

o°o

VÁSÁRHELYI, P.: Der Vergleich von Dokumentationssystemen

Die Arbeit der Western Reserve University am Gebiet des Vergleiches von verschiedenen Dokumentationssystemen unter streng bestimmten Umständen wird an Hand einer Studienreise in den Vereinigten Staaten behandelt. Die, in Zusammenhang mit dem Aslib-Cranfield-Test von CLEVERDON ausgearbeiteten Versuchsmethoden werden erörtert. Die Möglichkeit die Verschiedenheiten von Dokumentationssystemen mathematisch zu erfassen wird beschrieben.

oo°oo