Makay Árpád

Tezauruszok alkalmazása számitógépes információs rendszerekben

S z e g e d

1971.

B e v e z e t é s

A könyvtári és dokumentációs munka három fő területén kinálkozik alkalom a müveletek gépesitésére, automatizálására:

1. A publikálás müveletének automatizálását szolgálják a korszerü sokszorositó berendezések, lyukszalag-, mágnesszalag-szedőgépek. Felhasználhatók számitógépes rendszerek is: irógépből és vizuális megjelenitő készülékből álló, a számitógéppel on-line kapcsolatban lévő terminálokon gépelhető a publikáció. Az igy nyert információkat közvetlenül a számitógép szerkeszti, kiadványokat állit össze, esetleg vezérli a sokszorositó berendezést. A kiadványok automatikus előállitása mellett a rendszer másik előnye, hogy a publikáció számitógép számára közvetlenül hozzáférhető marad bármilyen egyéb feldolgozási célokra is.

2. A publikáció előállitásának automatizálása azt jelenti, hogy meglévő információk alapján a számitógép készit ujabb kiadványokat. Elsősorban bibliográfiák, kötetkatalógusok, névmutatók, tárgyszóindexek előállitására gondolhatunk. Ebbe a kategóriába sorolható az automatikus tartalmi kivonatok készitése is: a publikáció teljes szövegéből egy /általában lényegesen rövidebb/ szólánc előállitása, amely tartalmi szempontból a "lényeget", illetve a lényegre utaló kifejezéseket tartalmazza.

3. A három terület közül az információkeresés (használ-

latosabb terminológiával élve, az információvisszakeresés) automatizálása a legösszetettebb. Sajátos problémái mellett szoros kapcsolatai vannak az automatikus tartalmi kivonatolással, tárgyszó-indexeléssel.Lényegében arról van szó, hogy publikációk egy halmazából tartalmi vagy formai ismérvek alapján a számítógép válogasson és a kiválasztott publikációkat teljes egészükben vagy csak a rájuk vonatkozó utalások (azonosító szám, lelőhely, stb.) formájában az igénylő rendelkezésére bocsássa.

A tudományos és technikai publikációk mennyiségi növekedése indokolja az információkeresés automatizálásának szükségességét. Másrészt az információkeresés ma már főként tartalmi szempontok alapján elképzelhető, ezért tárolni csak tartalmilag feltárt és tartalmi ismérvek alapján könnyen hozzáférhető dokumentumot érdemes. Az információszolgáltatásnak pedig egyéni igényeket kell tudni kielégiteni, szemben a bibliográfiák, katalógusok, indexek (általában a kiadványok) rendszerével, amelyeket tömegkommunikációs eszközöknek tekinthetünk.

A publikációk tartalmi feltárásának /22/ müvelete, sőt magának a fogalomnak a jelentése is igen összetett. A publikáció egésze, mint információhordozó, bármilyen átdolgozás, kivonatolás, rövidités folyamán veszit információ-tartalmából. Mégis lehetnek olyan szempontjaink, amelyek alapján a publikáció által szolgáltatott információk egy részét lényegtelennek, más részét pedig lényegesnek tekintjük. Az utóbbi rész valamilyen formában történő megadását nevezzük tartalmi feltárásnak.

Információkereső rendszerekben a "lényeg" megadásának szempontjait a következőképpen jellemezhetnénk. Adva van bizonyos tipusu publikációknak (pl. számitástechnikával kapcsolatos folyóirat-cikkeknek) egy halmaza. Ugyanakkor elég jól körülhatárolható az az embercsoport (a felhasználói kör), amelynek tagjait érdekli a publikációk tartalma. Vizsgáljuk meg, milyen "szokásaik" vannak a felhasználóknak akkor, amikor valemilyen őket érdeklő téma után kutatnak a publikáció-halmazban. Ma a járható ut általában az, hogy néhány folyóiratot időrendben visszafelé lapozva átfutjuk a cimeket, összefoglalásokat, igy egy idő mulva összegyülik a számunkra elegendő "olvasnivaló". Ha még a már megtalált publikációk irodalomjegyzékét is figyelembe vesszük (azaz felhasználjuk a cikk irója korábbi irodalom-kutatásának eredményét), meggyorsithatjuk a munkát. Nem sok a munkára forditandó időnyereség akkor sem, ha a könyvtár állományába tartozó szakember végzi ezt helyettünk: megjelenik azonban a probléma, hogyan jelöljük meg számára a minket érdeklő témát, hogy valóban azt kapjuk, amire szükségünk van. Ugyanez a nehézség merül fel akkor is, ha a keresést számitógép végzi.

Nyilvánvalóan fel kell sorolni a keresett témára jellemző ismérveket: ezek általában a témával kapcsolatos fogalmakat jelölő szavak, kifejezések (kulcsszavak, deszkriptorok). Kérhetjük ezután azokat a publikációkat, amelyek két vagy több fogalom mindegyikével foglalkoznak; vagy egy fogalommal kapcsolatosak, de egy másikkal nem; stb.

A keresés folyamán most már az elsőrendü feladat, hogy megállapitsuk: egy adott publikáció kapcsolatos-e egy adott kifejezéssel megjelölt fogalommal. Nyilvánvalóan minden egyes felhasználó megjelenésekor végigolvasni minden egyes publikációt a publikáció-halmazból, és igy megállapitani a megfelelőségét, igen időigényes volna (emberi erővel egyáltalán lehetetlen). A tartalmi feltárás feladata információkereső rendszerben tehát éppen az, hogy hozzárendelje a publikációhoz mindazokat a fogalmakat, amelyek alapján a felhasználói kör bármelyik tagja később keresheti az adott publikációt. Számitógépes rendszerben ez még további feltételeket is szab: a fogalmakat kifejezésekkel irjuk le, igy a publikációhoz azokat a kifejezéseket (deszkriptorokat) kell rendelni, amelyek alapján a felhasználók keresni fogják. A későbbiekben a tartalmi feltárás eredményének mindig egy kifejezés-sorozatot (deszkriptor-sorozatot) tekintünk.

Az alábbiakban információvisszakereső számitógépi rendszerek két aktuális problémájának megoldására szolgáló, listaeljárásokkal megvalósitható kérdéséről lesz szó. Az egyik a dokumentumok tartalmi feltárásának (legalábbis részben) automatizálását tüzi ki célul, mig a másik az információs rendszerben tárolt, tartalmilag feltárt dokumentum-halmaz igényelt (speciális tartalmu) részének megjelölését (az igény megfogalmazását) célozza.

A dokumentumok tartalmi feldolgozásának teljes automatizálását részben a természetes nyelvekre vonatkozó grammatikai és szemantikai ismereteink jelenlegi hiányosságai, részben a dokumentumok terjedelme (ami a számitógép számára hozzá-

férhető adathordozóra való konvertálásának időigényességét jelenti) akadályozzák. A "tartalmi kivonatolásban" ezért részmegoldásokra szoritkozunk. Még sokáig szakemberre lesz szükségünk, aki a dokumentumot tartalmilag értékeli, jellemzőit összegyüjti. Egyszerüsithető azonban a forma, amelyben a számitógéppel az ismérveket közli. Az ismertetésre kerülő módszer egyik előnye pl. a következőkben foglalható össze. A dokumentum feldolgozását végző szakember a tartalmat szavakkal, szócsoportokkal (tárgyszavak ill. deszkriptorok) irja le. Ezek közismert, de általában az adott fogalomra nem kizárólagosan használt kifejezések: velük párhuzamosan élnek a fogalom egyéb megjelölései is. Ha a dokumentum csupán ezekkel a deszkriptorokkal kerül az információs rendszerbe, akkor a felhasználó is csak ezeket használhatja igényének megfogalmazására, másként nem juthat el az adott dokumentumhoz. A tartalmi feltárást végző ill. a felhasználó szakembernek, "egy nyelvet" kell tehát használni. Ezt segiti elő az un. tezaurusz: egy adott szakterület releváns kifejezéseinek gyüjteménye, mindegyik mellett egyebek között jelölve, hogy az illető kifejezés használható-e deszkriptorként, illetve ha nem, mit kell helyette használni. Kézenfekvő, hogy a tezaurusz alapján a számitógép hozza közös nevezőre a feldolgozót és a felhasználót, egyszerüsitve a tartalmi feltárás és az igények megfogalmazásának munkáját.

A tezaurusz használatát még egy lényeges szempont indokolja. A tartalmi feltárás során a dokumentumhoz rendelhetők olyan kifejezések, amelyek tulságosan "szük", speciális

fogalmakat jelölnek: ekkor az adott kifejezés mellé fel kell venni egy vagy több olyan deszkriptort, amely bővebb kategóriát jelent. A visszakeresés igy történhet akár a szükebb, akár a bővebb kategóriák alapján. Természetesen ehhez az szükséges, hogy a tezaurusz egy adott kifejezés mellett felsorolja a genetikus fogalmakat is.

Ez a "homogenizáló" fázis akkor sem hagyható el, ha eljutunk a tartalmi feldolgozás teljes automatizálásáig. A dokumentum szövegében illetve a kivonatban a tartalomra nézve releváns kifejezések (a gépi kivonatolás eddigi tapasztalatai szerint) nagy számát mindenképpen csökkenteni kell egy információvisszakereső rendszerben. A dokumentum és az igény összehasonlitásához szükséges idő csökkentésének egyik módja, (különösen nagy dokumentum-file-ok esetében) hogy elkerüljük a dokumentumhoz rendelt deszkriptorokban a redundanciát. Ugyanakkor a tartalmi feldolgozás eredményének leirásában, illetve az igények megfogalmazásában egyre nagyobb szabadságot kell adnunk, ami egyben azt is jelenti, hogy közelithetünk a természetes nyelvekhez.

1. Információs rendszerek

1.1. Az alábbiakban dokumentációs célokat szolgáló információtároló és visszakereső rendszerrel foglalkozunk. A rendszert "adatbanknak" is szokás nevezni, mivel nem szoritkozik a könyvtári gyakorlatban ismert dokumentációra: formai szempontból teljesen azonos rendszer megvalósithat személynyilvántartást, szabadalmi, müszaki konstrukciók nyilvántartását, stb. Ezért a továbbiakban a dokumentum terminológia használatakor nem szoritkozunk ennek csupán könyv, folyóirat-cikk, publikáció jelentésére, hanem egészen általánosan egy (legtöbbször szövegesen megfogalmazott) információ-hordozó egységet értünk alatta.

A dokumentum tartalmi vagy formai ismérveit deszkriptoroknak nevezzük. Ezek általában szavak, szócsoportok, számok, olyan egységek, amelyek a dokumentumra vonatkozó valamilyen tulajdonságot fejeznek ki. Közismert deszkriptorok pl. a könyvtári gyakorlatban a folyóirat-cikk tartalmi megjelölésére az ETO-számok, egy-egy szakterületre jellemző (tezaurusz-ban lefektetett) tárgyszavak, kifejezések; személynyilvántartásra gondolva a hajszin megjelölése, testmagasság, iskolai végzettség, stb. Egy dokumentumhoz több deszkriptor is tartozhat, ezeket egymástól függetlenül rendeljük a dokumentumhoz.

Egy információs rendszerben különböző tipusu deszkriptorok szerepelhetnek. Némelyek a tartalom, mások formai jegyek leirását szolgálják. A deszkriptor indikátora jelzi,

hogy az általa képviselt tulajdonság a dokumentumot mely oldalról közeliti meg. Az indikátor lehet valamilyen azonositó (pl. NÉV, ETOSZÁM,...), de gyakrabban számokkal jelezzük (legalábbis a számitógép belső ábrázolási formájában). Az indikátorok rendszere utal az információs rendszer felhasználási körére: a dokumentumokról tárolt ismérvek jellegét illetve a visszakeresés szempontjait határozza meg.

Az alábbi példákban az indikátort a deszkriptor szövegétől / jel választja el a C indikátor pedig a tartalmi ismérvet jelöli:

SZERZŐ/Weissmann,G.
C/programozási nyelvek
C/LISP
C/szövegfeldolgozás számitógéppel
C/listák

A formai ismérvek (szerző, kiadó, leltári szám, stb.) alapján történő keresés nem okoz olyan jellegü problémákat, amellyel a dolgozat foglalkozik. A továbbiakban, amikor deszkriptorokról beszélünk, gondoljunk mindig a publikáció tartalmi feltárása során nyert kulcsszavakra, tárgyszavakra. A könyvtári és dokumentációs gyakorlatban is ez az általánosan elfogadott jelentése.

Bár a tartalmi feltárás során a dokumentumhoz rendelt deszkriptorokat egymástól függetleneknek tekintjük, az információkereső rendszerben használható, illetve használt deszkriptorok rendszere strukturális tulajdonságokkal rendelkezhet: elemeit többféle kapcsolat füzheti össze. A genetikus és specifikus fogalmakat jelölő kifejezések közötti

reláció például a deszkriptor-halmaz parciális rendezését adja; a bevezetőben említett "kitüntetett deszkriptorok" a struktura ekvivalencia-osztályozását jelentik, stb. A deszkriptorrendszer strukturája alkalmas arra, hogy 1./ a publikáció tartalmi feltárása során kapott deszkriptor-sorozatban és 2./ a felhasználói igényben szereplő deszkriptorokon a strukturában megengedett müveleteket végezzünk.

A deszkriptor-halmaz elemeinek felsorolására (véges, de az időben változó, bővülő halmaz) és az elemek közötti kapcsolatok jelölésére alkalmas a tezaurusz. A tezauruszok közös vonása, hogy tartalmazzák a tudományok vagy a technika egy-egy területén előforduló fogalmakra a szakirodalomban általában használt kifejezéseket.

A tezaurusznak, mint halmaznak az elemeit nevezzük kulcsszavaknak. A legelterjedtebb strukturális tulajdonságok, amelyeket tezauruszokban szokásos megadni, a következők:

a.) Az ekvivalens kifejezések (a kifejezésekkel leirt fogalmak azonosak, illetve az információkeresés szempontjából azonos jelentésüek) rendszerének leirása. A tezaurusz minden egyes kulcsszóhoz az ekvivalens kifejezés-osztály egy kitüntetett elemét rendeli. Az osztály kitüntetett elemét megengedett deszkriptornak, az osztály többi elemét tiltott deszkriptornak nevezzük. Szokás deszkriptornak illetve nem-deszkriptornak is nevezni.

b.) A fogalom-hierarchiák jelölése. Ehhez elegendő a kitüntetett deszkriptoroknak megfelelő kulcsszavak mellett felsorolni az általa képviselt fogalom genetikus, illetve

specifikus fogalmait. Szokásos még rokon-fogalmak megjelölése is. Természetesen problématikus, hogy a tezaurusz milyen mértékben általánositson, illetve specializáljon. Információkereső rendszerben pl. a tul általános fogalom a publikációk közül sokra érvényes, igy a visszakeresési eljárás "bőbeszédü" lesz. A fogalom-hierarchia felépitése az adott területen igen nagy szakértelmet követel, ezenkivül figyelembe kell venni a tezaurusz felhasználási területét is. Általában csak számitógép képes megállapitani, hogy a hierarchia valóban parciális rendezettség-e, azaz vannak-e a tezaurusz-ban ellentmondások.

c.) A tezaurusz tartalmazhat a kulcsszavakra vonatkozó statisztikai paramétereket. Elsősorban egy publikációhalmaz elemeinek (automatikus vagy szakemberek által végzett) tartalmi kivonatolása során nyert összes deszkriptoron belüli gyakoriság jegyezhető fel. Megállapitható a kulcsszó előfordulásának gyakorisága az információvisszakereső rendszer igényeiben, stb. Ezek a paraméterek segitséget nyujthatnak az automatikus tartalmi kivonatolás egyes fázisaiban, de szükségesek a tezaurusz állandó korszerüsitéséhez, bővitéséhez, illetve az adott rendszerhez történő "hozzáidomitásához".

A dokumentum egészének (teljes szövegének) tárolása az információs rendszerben általában terjedelme miatt nem megvalósitható, bár pl. mikrofilm-tárolóban nem lehetetlen. Továbbá a dokumentumok keresése a deszkriptorok alapján történik, igy a visszakeresés szempontjából csupán a dokumentumok azonositását szolgáló ismérvek szükségesek: leltári

szám, szerző, cim, kiadási adatok, lelőhely. Sokszor ezekhez még egy természetes nyelven megfogalmazott tartalmi kivonat kapcsolódik, amely a felhasználó számára nyujt információkat, még mielőtt a publikációt kézbe venné.

A dokumentum azonositását szolgáló ismérveket, együtt a tartalmi feltárás során kapott deszkriptorsorozattal nevezzük dokumentum-rekordnak, rövidebb terminológiával az információs rendszer dokumentumának. Az információs rendszerbe felvett dokumentum-rekordok összességét dokumentum-file-nak nevezzük. A dokumentum-file a rendszer üzemelése, használata közben uj dokumentum-rekordok belépésével állandóan bővül.

1.2. Az információs rendszereket, céljukat tekintve két csoportba szokták sorolni:

a.) Időszakonként tájékoztatást kivén nyujtani a dokumentum-file tartalmáról. A dokumentumokat egy vagy több szempont szerint sorbarendezve, lista formájában megjelenik a file. Tájékoztató kiadványok, bibliográfiák, rewiev-k, indexek formájában publikálhatók. A deszkriptoroknak általában a rendezési elvekben van szerepük. A rendszerben készülő kiadványok felhasználók kisebb-nagyobb csoportja számára készülnek, tömegkommunikációs eszközök.

b.) A több ezer vagy százezer dokumentumból bizonyos tulajdonságuak egy-egy csoportját bármikor szolgáltatni tudja. A tulajdonságok leirására a deszkriptorokon kivül általában logikai müveleteket használhatunk. A logikai müveletek (és, vagy, negáció) szerepe az, hogy deszkriptorok együttes előfordulását vagy hiányát irhassuk elő. A tartalmi ismérvek

megadásának tipikus formáját a következő példa illusztrálja:

Dokumentáció és információvisszakeresés vagy
automatikus információvisszakeresés vagy
információvisszakereső számitógépes rendszerek és nem IBM

A tulajdonságok közlését a számitógéppel (felhasználói) igénynek nevezzük, azt az eljárást pedig, amely a dokumentum-file-ból a megfelelő csoportot meghatározza és a felhasználó számára hozzáférhetővé teszi, információvisszakeresésnek. A rendszer által szolgáltatott információk egyénnek vagy szük csoportnak szólnak.

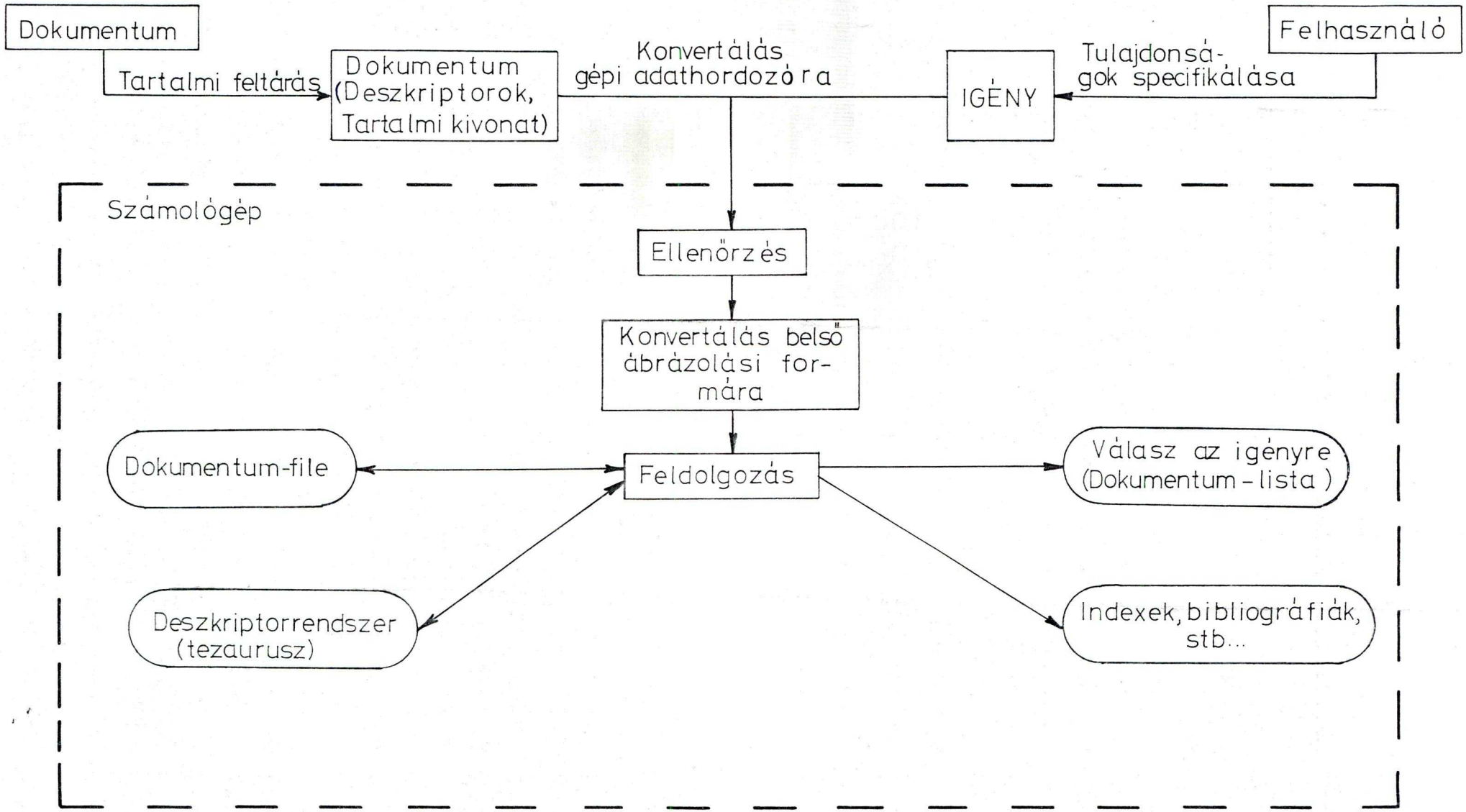
A két cél egymástól mereven nem határolható el, minden esetre a visszakereső rendszert a "kérdés-válasz" jellemzi. Az igények lehetnek hosszabb időszakra nézve állandóak vagy esetenkéntiek, ebből a szempontból a felhasználó (vagy a felhasználó és a gép között közvetitő személy) off-line vagy on-line (konverzációs) kapcsolatban lehet a rendszerrel. Másrészt a válaszadásra fordítható idő is igen változó (a könyvtári kölcsönzésben percek alatt, mig a témafigyelő szolgálatban havonként egyszer kell kielégiteni az igényeket). Mindezek a körülmények befolyásolják

- a dokumentum-rekordok és a dokumentum-file strukturáját
- a dokumentum-file gépi adathordozójának kiválasztását és alkalmazhatóságát (mágnesszalag, mágneslemez, mikro-film, stb.)
- a visszakereső eljárás felépitését
- használható számitógép hardware-strukturáját és konfigurációszükségletét.

1.3. Az 1. ábra az információs rendszerek általános strukturáját szemlélteti. Az egyedi rendszerek a különböző blokkok tényleges tevékenységeiben térnek el egymástól.

A dokumentum, rendszerbe lépésekor először tartalmának megfelelő szakember kezébe kerül, aki a tartalmi feldolgozást végzi. A számitógépi információs rendszerek használatában a leggyakoribb kifogás, hogy ez a fázis (legalábbis jelenleg) elkerülhetetlen. Az automatikus kivonatolással kapcsolatos kutatások (bár széles körben folynak) eddig nem eredményeztek kielégitő megoldást. Ez elsősorban nyelvészeti problémák megoldatlanságának következménye, amihez hazánkban is még egy sulyos kérdés járul: többnyelvüség a szakirodalomban.

A feldolgozást végző szakember a dokumentumhoz tartalmi kivonatot készit, deszkriptorokat rendel hozzá, formai és egyéb ismérveit az indikátorrendszernek megfelelően összegyüjti. A számitógép ebben a fázisban is egyre inkább szerephez jut. A dokumentum azonositását szolgáló ismérvek, a tartalmi kivonat és a tartalmi feltárás eredményeként kapott deszkriptorsorozat gépi adathordozóra (lyukszalag, kártya, mágneskártya, stb.) kerül, amely már számitógép számára közvetlenül hozzáférhető. Függetlenül attól, hogy a deszkriptorrendszer rendelkezik-e valamilyen strukturával, mindenképpen formai ellenőrzéseket kell első lépésben a számitógépnek végrehajtania. A deszkriptor-szótárral (amely a visszakeresési eljárás végrehajtásához is általában szükséges, tehát rendelkezésre áll) összehasonlitásokat végez. Ha ezenkivül a deszkriptorrendszernek strukturája is van (azaz tezaurusz is ren-



1. ábra. Információs rendszer.

delkezésre áll), a struktura alapján egy standard eljárást hajt végre a dokumentumhoz rendelt deszkriptor-listán, amelynek eredményeképpen elmaradhatnak, megváltozhatnak, illetve bővülhetnek a lista elemei. Ezzel egyidőben jelzéseket ad a rendszer üzemeltetőjének, ami alapján korrigálható a belépő dokumentum-rekord illetve változtatható, bővithető a tezaurusz strukturája. A változtatások elsősorban a terminológia egységesitése miatt szükségesek. Nem elhanyagolható másik szempont, hogy az egyes fogalmakat jelölő kifejezések mellé felvegyük az adott fogalomra nézve genetikus fogalmakat jelölő kifejezéseket is. Igy biztositható, hogy a visszakeresési eljárás rátalál a dokumentumra akár a szükebb, akár a bővebb fogalom alapján.

A specifikus fogalmakkal ugyanez az eljárás már nem ajánlatos: a tágabb kategóriából a szükebbre segéd-információk nélkül nem következtethetünk.

A változatlanul hagyott vagy módositott dokumentum-rekordot, ezután a számitógép belső ábrázolási formára konvertálja, sorszámozva csatolja a dokumentum-file-hoz. A dokumentum-rekord strukturáját elsősorban a visszakeresési eljárás szempontjai határozzák meg: gyors hozzáférhetőség a deszkriptorokhoz, szövegszerü összehasonlitások egyszerü végrehajthatósága. Mind a rekord, mind a rekord mezői változó hosszuságuak, ezért általában szét kell választani az indikátort és a tulajdonságot jelölő karaktersorozatot illetve számot. A indikátor mellett jegyezzük fel a deszkriptor szövegére vonatkozó lényeges információkat (hossza, helye

a rekordon belül, stb.), igy a visszakeresési eljárásban sorrakerülő összehasonlitások nagy része fix hosszuságu egységeken végezhető. Bibliográfiai rekordokra az USA-beli MARC-II. szabvány rögzit ezeknek a kritériumoknak megfelelő mágnesszalag-tárolási formát /3/.

A felhasználó a dokumentum-file anyagához az igény közlésével fér hozzá. Az igény elemi egysége a deszkriptor, szerepe pedig az, hogy a dokumentum-halmazt két részhalmazra osztja: az egyik csoport dokumentumaihoz hozzá van rendelve, a másik csoport dokumentumaiban nem szerepel az illető deszkriptor. Az igényben a két részhalmaz közül bármelyiket kijelölhetjük, ezt a részhalmazt azután további deszkriptorok segitségével ujra részhalmazokra bonthatjuk, ilymódon az igény a dokumentum-file egy meghatározott részét választja ki. A használatos információs rendszerek a vázolt metszetképzés müvelete mellett általában az egyesitést is megengedik, de ennél többet csak akkor adhatunk, ha a deszkriptor-rendszert strukturával látjuk el.

1.4. A <u>visszakeresési eljárás</u> tényleges <u>végrehajtására</u> alapvetően két mód kinálkozik:

a./ A dokumentum-file minden egyes elemét összehasonlitjuk az igénnyel, a megfelelőeket különválasztva a felhasználó rendelkezésére bocsátjuk. A dokumentum-rekord és az igény összehasonlitása időigényes eljárás, nagy dokumentum-file-ok esetében igy ez az ut nem járható.

b./ A dokumentum-rekordokat deszkriptoraiknak megfelelő sorrendben tároljuk. Ha egy dokumentumhoz több deszkriptor tartozik, a dokumentum-rekord teljes egészében többször is-

métlődik. A deszkriptorok segitségével megfogalmazott igény bejelentésekor adott deszkriptor mellett megtaláljuk az igényelt dokumentumokat is. Az ismétlődéseken kivül a módszer további hátránya, hogy pl. két deszkriptor együttes jelenlétének ellenőrzése igen nehézkes.

A két (szekvenciális és invertált) módszert egyesiti a kombinált információvisszakeresési eljárás /23, 25, 8/. Amikor a dokumentum a rendszerbe lép, egy sorszámot (indexet) kap. A dokumentum-file-ba a rekordok sorszám szerint növekvő rendben épülnek be. Ezzel egyidejüleg módosul az a deszkriptor-file, amelyben minden egyes deszkriptor mellett fel vannak sorolva azoknak a dokumentumoknak az indexei, amelyekhez az illető deszkriptort hozzárendeltük. Ha most egy igény megjelenik, a keresési eljárás azzal kezdődik, hogy

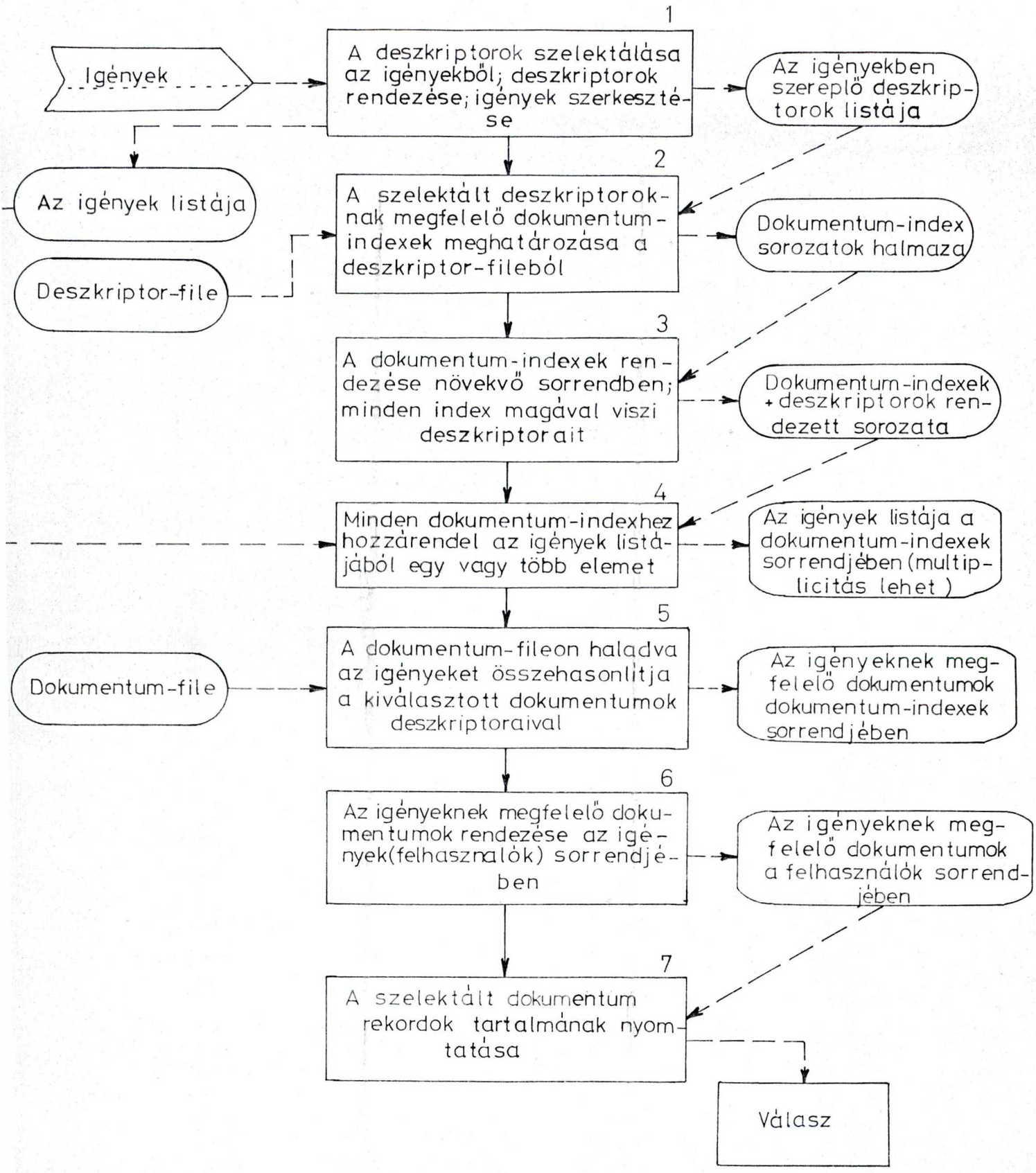
- szelektáljuk az igényből az összes előforduló deszkriptort
- a deszkriptor-file-ból kiolvassuk mindazokat az indexeket, amelyek a szelektált deszkriptorok mellett találhatók.

Ezután a tényleges összehasonlitást csak azokon a dokumentumrekordokon hajtjuk végre, amelyek indexeit kigyüjtöttük.

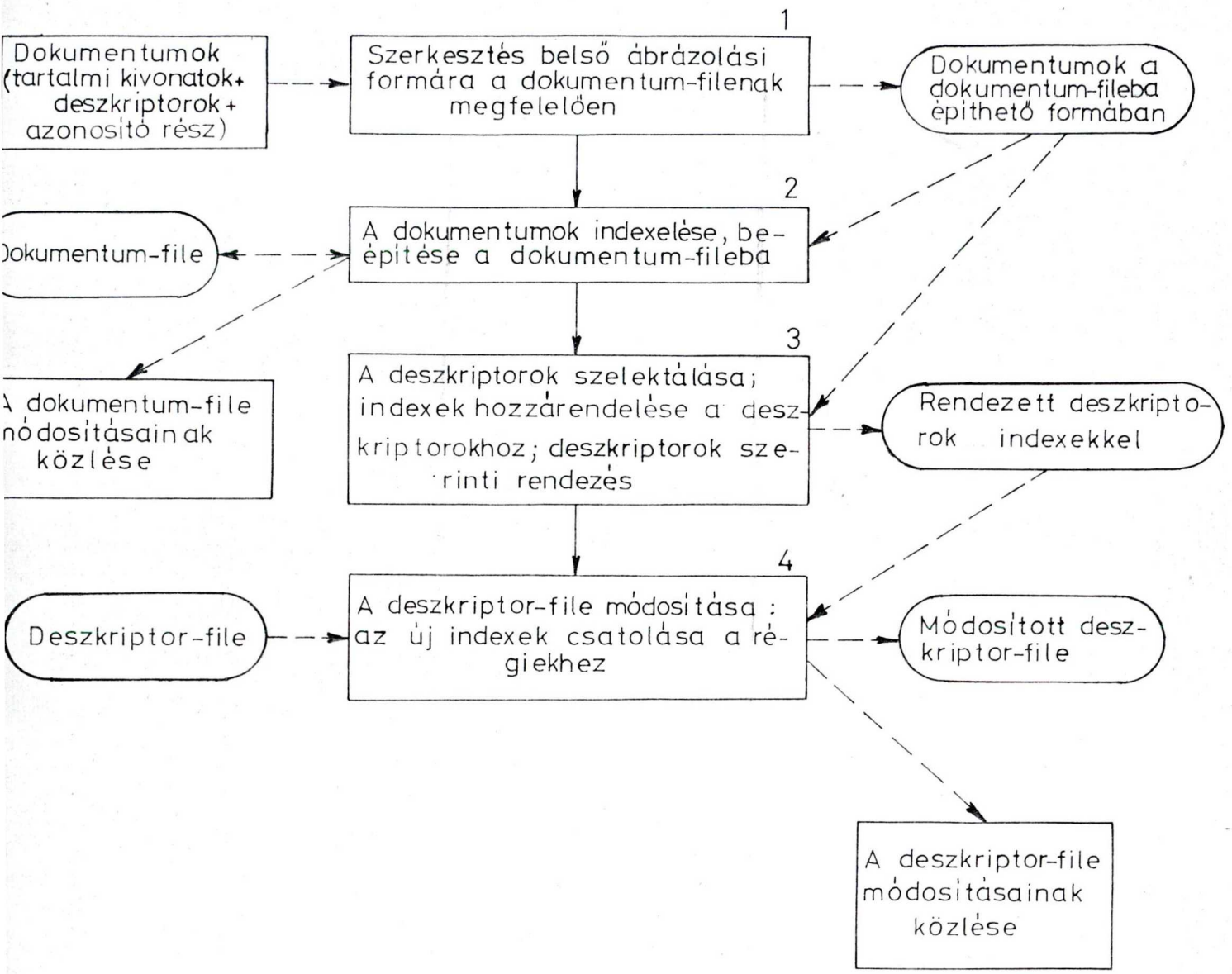
Az egy igényre forditott keresési időt lényegesen csökkenti az igények csoportositása. A dokumentum-file egyszeri végigfutása közben több felhasználó igénye elégithető ki. A 2. és 3. ábra kombinált információvisszakeresési eljárást alkalmazó rendszerben szemlélteti a dokumentum belépését illetve a visszakeresés végrehajtását. A téglalapok az egyes

tevékenységeket jelölik, közöttük a folyamatos vonallal jelölt nyilak a tevékenységek sorrendjét jelzik. A többi blokk a kiinduló, a közbülső adatokat, illetve az eredményeket jelölik. A szaggatott vonallal jelölt nyilak az egyes tevékenységek input-adataiból illetve outputjaiba mutatnak.

A továbbiakban felépitjük a deszkriptorrendszer strukturájának, a tezaurusznak lista-modelljét, definiáljuk azokat a müveleteket, amelyek alkalmasak e strukturának birtokában kiválasztani az igénynek megfelelő dokumentum-halmazt.



2. ábra. Kombinált információvisszakeresési eljárás



3. ábra. Az információs rendszer bővítése

2. Listák és elemi müveletek

2.1. A számitógépet hagyományosan numerikus jellegü feladatok megoldására használják. Az eközben felhasznált algoritmusok (bizonyos pontossággal megadott) racionális számokon végeznek aritmetikai müveleteket, számokkal képeznek relációkat, számokból épitenek fel vektorokat, mátrixokat. Számok tárolására általában egy-egy memóriabeli rekesz (byte-szervezésü gépeken byte-ok egymásutáni sorozata) szolgál, mig a tömböket a memória egy intervalluma képviseli. Ez az ábrázolási forma két speciális tulajdonsággal rendelkezik:

a./ Egy szám ábrázolására a memóriában mindig ugyanannyi, illetve előre meghatározott nagyságu hely szükséges, igy a számokból képzett tömbök memóriabeli helyigényének megállapitásához elegendő az elemek számának ismerete.

b./ A tömb egy elemére egyértelmüen hivatkozhatunk a tömböt tároló memória-intervallum kezdőcimével és az elem tömbön belüli sorszámával.

Mindkét előnyös tulajdonság annak köszönhető, hogy az alapmennyiségeket (a számokat) ismert hosszuságu memóriaegység tárolja (szóban, byte-ban vagy bit-ben számolva). A nem-numerikus jellegü feladatok éppen ezt az alapfeltételt nem teljesitik: az elemi mennyiségek (karaktersorozatok, pl. szavak; decimálisan ábrázolt számok) változó hosszuságuak. Másszóval mindaddig nem ismerjük az adat tárolásához szükséges memória méretét, amig az adat ténylegesen rendel-

kezésre nem áll. Ebből következik uj módszerek alkalmazásának szükségessége a memória optimális kihasználása végett.

A probléma megoldásához jelentősen hozzájárult a byte-szervezésü számitógépek megjelenése: a vezérlő illetve aritmetikai egység képes változó (az utasitásban megadott vagy szó-határjellel definiált) hosszuságu alapegységeket értelmezni. Másrészt egy-egy feladattipus kapcsán különböző számitástechnikai eljárások (verem-memória, dinamikus cim-kezelés, stb.) alakultak ki.

Az alap-mennyiségek változó hossza az adatok struktúrájának számitógépi ábrázolásában is ujat jelentett. Ha pl. egy vektor komponenseit továbbra is indexüknek megfelelő sorrendben egymás utáni memória-cimeken helyezzük el, akkor egy adott elem cime az összes megelőző elemek hosszának függvénye. Ezenkivül egy vektorelem helyettesitése a korábbinál hosszabb karaktersorozattal együtt jár azzal, hogy az összes nagyobb indexü komponenseket a memóriában feljebb kell tolni.

Egyébként sem csupán arra van szükség, hogy a meglévő numerikus eljárások adatstruktúráját másoljuk uj tartalommal. Ahogy szélesedik a számitógépek felhasználási köre, egyre ujabb strukturák gépi ábrázolására van szükség, ami együtt jár az algoritmusok elemi müveleteinek változásával.

2.2. A <u>listák</u> /2, 7, 19, 27/ a gráfelméletből ismert irányitott bináris (dichotomikus)gráfok strukturájával irják le az adatokat. A gráf pontjai a <u>cellák</u> illetve <u>atomok</u>. Az atomok képviselik az elemi adattipusokat, amelyek általá-

ban karaktersorozatok. Az atomokból élek nem indulnak ki, mig minden cella a gráf két pontjára mutató élt reprezentál. A listamüveletek egy része uj listákat állit elő vagy meglévőket módosit, mig más része egy cellából egy másik cellába vagy atomba vezető utat definiál, ezáltal allistát (speciális esetben atomot) jelölve ki.

Szükségesek, de nem takinthetők listamüveleteknek azok a müveletek, amelyeket atomokon definiálunk és az atomok belső strukturáján alapulnak (aritmetikai, logikai müveletek, az atom hosszának meghatározása karakterben számolva, két elem összehasonlitása, stb.).

A továbbiakban szereplő példákat az Országos Müszaki Könyvtár és Dokumentációs Központ által 1970-ben közreadott Általános müszaki fogalmak tezaurusza /1/ szolgáltatja. Az alapanyag előállitásában (a rendelkezésre álló angol nyelvü tezaurusz átrendezésében a magyarra forditás után, a három rész rendezésében, a permutációs fejezet kidolgozásában) a Kibernetikai Laboratórium Minszk-22 számitógépe is szerephez jutott. Az Általános müszaki fogalmak tezaurusza megkülönböztet deszkriptoroknak és nem-deszkriptoroknak nevezett kulcsszavakat. A nem-deszkriptorok mellett megtalálhatjuk az ugyanazt a fogalmat kifejező deszkriptort. A deszkriptoroknál megjelöli a rokon fogalmakat, a specifikus fogalmakat és a genetikus fogalmakat reprezentáló deszkriptorokat. Az alábbi lista-modell tezaurusz-reprezentációjában az atomoknak a kulcsszavak (deszkriptorok és nem-deszkriptorok) felelnek meg, a listák strukturái pedig a deszkriptorok fogalom-kapcsolatait fejezik ki. Pontosabban szól-

va: minden kulcsszónak megfelel (egyértelmüen, de nem kölcsönösen egyértelmüen) egy lista, amelynek speciális al-listáiban szereplő atomok képviselik az azonos jelentésü deszkriptort, a speciális, genetikus és rokon fogalmak halmazait.

A müveletek jelölési módjában a LISP /29/ programozási nyelv koncepciója az irányadó.

2.3. A listák elemi egysége az atom. Megjelenési formája a karaktersorozat, hossza nem korlátozott. A listaműveletekben mint oszthatatlan egység vesz részt. Pl.

Fourier-sorok

integrálok

-512

A4

Speciális (nem lista-) müveletek, amelyeket atomokon definiálunk, tulajdonithatnak az atomoknak "belső" strukturát. Ez alapján osztályozhatjuk az atomokat: számok, szavak, kifejezések, logikai értékek, kódok, stb. Deszkriptorokra gondolva lehet az indikátor és a tulajdonságot kifejező karaktersorozat egysége /T: nem-deszkriptor vagy tiltott kifejezés, D: deszkriptor/:

T/ harmonikus analizis

D/ Fourier-analizis

Hogy egyetlen atomból is lehessen listát felépiteni, illetve, hogy a cellákból mindig két él fusson ki, bevezetjük a speciális üres atom fogalmát, amelyet O jelöl.

2.4. Cellának nevezzük az (x,y) párt, ahol x is és

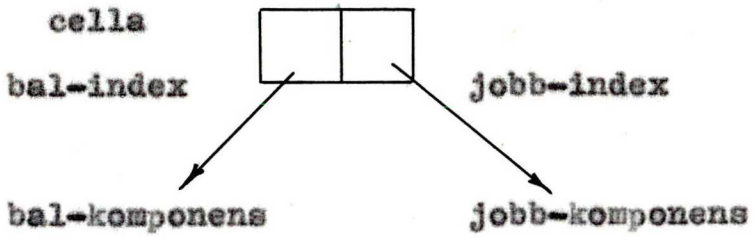
y is atom vagy cella lehet. x a bal-, y a jobb-komponens. Pl:

(T/ harmonikus analizis, D/ Fourier-analizis)

(D/ Fourier-analizis, 0)

(T/ harmonikus analizis, (D/ Fourier-analizis, 0)).

A cellákat irásban a fenti módon zárójelpárral és vesszővel jelöljük. A grafikus ábrázolás részben szemléletes, másrészt a cella számitógépi reprezentációjához is közel áll: a cella bal- és jobb-indexből áll, amelyek rendre a bal- illetve jobb-komponensre mutatnak. Számitógépben az indexek olyan cimek, amelyek a komponensek memóriabeli helyére utalnak.



2.5. Listának nevezünk

1./ Egy atomot, vagy

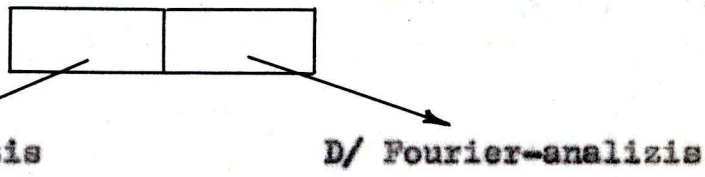
2./ listákból képzett párt (cellát).

Irásban a lista minden egyes cellájára egy zárójelpár utal. A listát egyértelmüen megjelölhetjük (azonosithatjuk) a legkülső zárójelpárnak megfelelő cellával, a lista-fejjel. A lista bármelyik cellájához (zárójelpárhoz) illetve atomjához tartozik egy al-lista. Az al-lista fejét (azaz magát az al-listát) a listában definiált uttal adhatjuk meg: az a cellából (a lista feje) a b cellába vagy atomba vezető uton értjük az $a_1,a_2,\ldots,a_n (n>0)$ sorozatot, ahol $a_1=a, a_n=b$

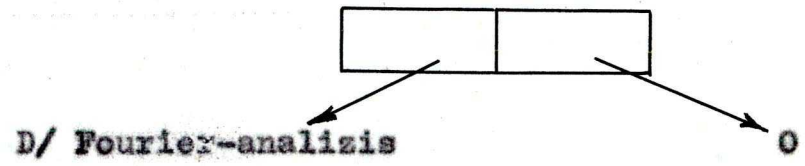
és minden a_i (i = 1,2,...,n-1) olyan cella, amelynek a_{i+1} vagy bal-, vagy jobb-komponense.

Példák:

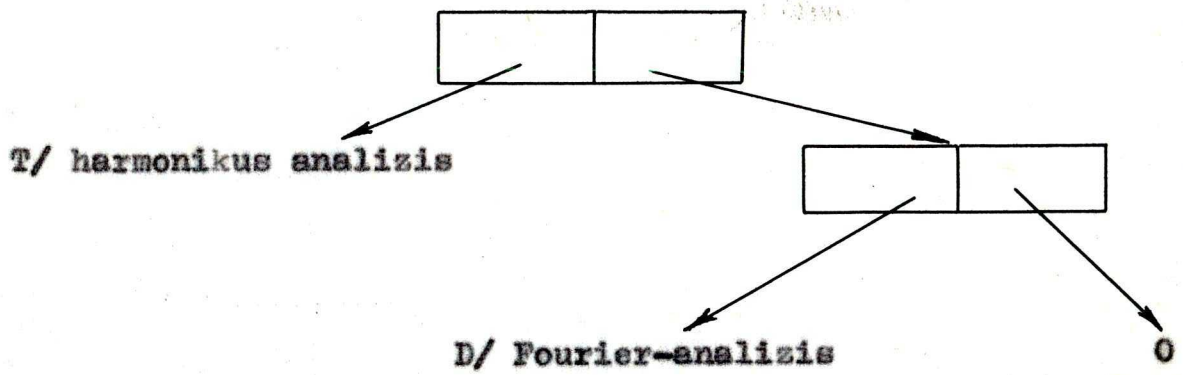
1./ (T/ harmonikus analizis, D/ Fourier-analizis)



2./ D/ Fourier-analizis, 0)



3./ (T/ harmonikus analizis, (D/ Fourier-analizis, 0))



2.6. L_1 és L_2 lista. Az (L_1,L_2) pár (cella) egy uj lista feje, azaz a pár-képzés kétváltozós, listákon értelmezett és lista-értékü müvelet. Hagyományos müveleti jelek hiányában függvény-jelölést használunk:

$$(L_1,L_2) = \text{cons}(L_1,L_2)$$

Számitógépben ez egyetlen uj cella képzését jelenti, amelynek bal-indexe az L_1, jobb-indexe az L_2 lista fejére mutat.

A függvény-jelölésben a zárójelpárnak és vesszőnek

argumentum-kijelölő, mig a cella-irásmódban bal- és jobb-komponenset meghatározó szerepe van. Ezt a kettősséget az egyes programozási nyelvek ugy oldják meg, hogy valamelyik jelölésmódot kizárják. A fordító- illetve értelmező-programok szempontjából a függvény-jelölés kedvezőbb: balról haladva az argumentum(ok) olvasásakor a végrehajtandó művelet már rendelkezésre áll.

2.7. Al-lista illetve ut definiálására szolgál a hd és tl művelet. Mindkettő nem egyetlen atomból álló L listán van értelmezve. hd(L) az L lista fejének bal-komponense (cella vagy atom). tl(L) az L lista fejének jobb-komponense (cella vagy atom).

A cons művelettel ellentétben a hd(L) és tl(L) nem képez uj cellát, hanem kijelöl egy meglévőt.

2.8. A cons, hd, tl és egyéb elemi műveletekből kompozició-képzéssel alkothatunk összetett kifejezéseket. Néhány egyszerü összefüggést találunk a definiált műveletek között:

$$hd(cons(L_1,L_2)) = L_1$$
$$tl(cons(L_1,L_2)) = L_2$$
$$hd(hd(cons(L_1,L_2))) = hd(L_1)$$

- - - - - - - - - - - - - - - - -

Viszont

$$f(L) = cons(hd(L),tl(L)) \neq L,$$

mivel a cons művelet mindig egy uj cellát definiál, Az f(L) lista strukturája azonos L strukturájával, de lista-fejük különböző cella.

2.9. Legyen

$$H = (L_1, L_2, \ldots, L_n)$$

listák egy rendszere. (Rendszert kell mondanunk halmaz helyett, mert elemeit szoros kapcsolatok füzhetik össze. Különböző listáknak lehetnek közös al-listái; pl. az L_i és $L_j = \mathrm{cons}(\mathrm{hd}(L_i), \mathrm{tl}(L_i))$ listák csak a lista-fejben különböznek. Másrészt minden cella illetve atom képviselhet egy listát: közülük csak bizonyosakat soroltunk H-ba, a többit al-listaként kezeljük.)

A cons, hd, tl müveleteket (ismételten) alkalmazva a H rendszer elemeire, olyan H' rendszert kapunk, amelynek H része: $H \subseteq H'$. Nem változtattunk meg egyetlen L_i listát (sőt cellát) sem.

A set, sethd, settl müveletek cellákat módositanak. A $\mathrm{set}(L_1, L_2)$ függvény alkalmazásának eredményeképpen az L_1 cella helyébe L_2 kerül. A $\mathrm{sethd}(L_1, L_2)$ illetve a $\mathrm{settl}(L_1, L_2)$ az L_1 cella bal- illetve jobb-komponense helyébe az L_2-t helyettesiti.

Ilymódon a set müveletek azt, vagy azokat a listákat, amelyekből való cellát módositanak, megváltoztatják.

A listák számitógépi tárolásában a cella-módositó müveletek egy igen lényeges problémát vetnek fel. Az atomoktól eltekintve a listák elemei (a cellák) fix hosszuságu memóriaegységekben (rekeszekben, byte-okban) tárolhatók. Maguk a listák viszont változó számu cellából épülnek fel, igy egy H listarendszer számára szükséges memóriaintervallum a rajta végzett müveletektől (a belőlük képzett listák hely-

igényétől) függ. Listákat ábrázoló programozási nyelvek, illetve programozási rendszerek ezért általában a lehetőségekhez (memóriakapacitás, felhasználó igény-jelzései, stb.) mérten maximális munkamezőt jelölnek ki a program számára, amelyben el kell férni mind a kiinduló adatokat képviselő, mind a generált litáknak. A listákat programozási nyelvekben a változóknak megfelelő azonositók jelölik, amelyeket a program irója deklarál. Minden azonositó tulajdonképpen egy lista-fej cime a munkamezőben. Az al-listáknak megfelelő celláknak nincs azonositójuk, ugyanigy nincs a kifejezések feldolgozása során keletkező közbülső listákat reprezentáló celláknak sem.

A munkamezőbe tehát egyre ujabb és ujabb cellák kerülnek. Ezzel egyidejüleg cellák válhatnak feleslegessé: a setmüveletek pl. "lekapcsolnak" al-listákat egy-egy listáról. A munkamező optimális kihasználása érdekében igy időnként a "felesleges" cellákat fel kell szabaditani, hogy helyüket ujra felhasználhassuk.

Tegyük fel, hogy az M munkamezőben az ($a_1,a_2,\ldots,a_n$ változókhoz rendelt) $L_1,L_2,\ldots,L_n$ listákat tároljuk. A c cellát az M mezőben foglaltnak nevezzük, ha az $L_1,L_2,\ldots,L_n$ listák valamelyikének lista-fejéből ut vezet c-hez.

Az M mező többi cellája "felesleges", mivel sem azonositóval, sem al-lista képzéssel nem tudunk rá hivatkozni, igy időnkénti (egy-egy kifejezés feldolgozása utáni) regeneráló eljárással felszabaditható további célokra.

A c cella foglaltsága függvénye az M mezőhöz rendelt változóknak. Minden a_i változóhoz (azonositóhoz) tar-

tozik egy L_i lista-fej. Asszociációs listának nevezzük az M mezőben az $(a_1,L_1),(a_2,L_2),\ldots,(a_n,L_n)$ sorozatot.

Az asszociációs listákat általában valóban lista-formában tárolják a programozási rendszerek, mivel a program futása során éppugy változhatnak, mint a kifejezések segitségével felépitett listák. (Deklarációk, blokkhatárok, eljáráshivások egyrészt az asszociációs lista hosszát változtatják, másrészt az azonositókhoz uj listákat rendelnek.)

2.10. Feltételes kifejezések képzéséhez logikai értékü függvények definiálására van szükség.

L,L_1,L_2 listák. Az atom (L) függvény értéke igaz, ha L egyetlen atom, különben hamis. Az egyenlőség értelmezése már problémákat okoz: definiálhatjuk a strukturák azonosságaként vagy egyszerüen a lista-fejek indexeinek megegyezése alapján. Ez utóbbi általában trivialitás, igy a következő definició hasznosabb:

a./ L_1,L_2 mindegyike atom. $eq(L_1,L_2)$ igaz, ha az L_1,L_2 atomok megegyeznek (mint karaktersorozatok).

b./ Ha L_1 és L_2 közül az egyik atom, a másik cella, akkor $eq(L_1,L_2)$ hamis.

c./ L_1,L_2 cellák, B_1,B_2 illetve J_1,J_2 a bal- illetve jobb-komponenseik. $eq(L_1,L_2) = eq(B_1,B_2) \wedge eq(J_1,J_2)$.

3. A tezauruszok listái

3.1. Információkereső rendszerünkben listákat használunk mind a tezaurusz strukturájának leírásában, mind pedig a felhasználói igények megfogalmazásában. Eközben a tezaurusz kulcsszavai (amelyek megegyeznek a dokumentumokhoz rendelhető deszkriptorokkal) kétféle szerepüek:

a./ A tezaurusz listáinak azonositói, azaz nevek. Minden egyes kulcsszóhoz tartozik a tezauruszban egy lista (több kulcsszóhoz esetleg ugyanaz a lista): a tezaurusz asszociációs listája feleíteti meg a kulcsszavaknak a listafejeket. A felhasználói igény alapegysége a lista-kifejezés, amelynek atomjai a kulcsszavak, szerepük pedig azonositó (változónév) jellegü. Az igény egy kulcsszavának értéke az a lista, amelyet a tezaurusz asszociációs listája a kulcsszóhoz rendel. A felhasználói igényekben megengedhetünk olyan kulcsszavakat is, amelyeket nem tartalmaz az asszociációs lista. A kulcsszónak, mint változónak az értéke legyen ilyenkor az egyetlen atomból (az adott kulcsszóból) álló lista pl. a tulajdonnevek. (IBM, VIDEOTON) nem vehetők fel a tezauruszba, illetve a nevekben nem érhetjük el még a viszonylagosan teljes felsorolást sem.

b./ A tezaurusz listáiban ugyanezek a kulcsszavak (de már csak a deszkriptorok) a listák atomjai lesznek, azaz elemi adatok, és nem azonsoitók.

Az információkereső rendszer tezaurusza tehát két részből áll: az asszociációs listából és a deszkriptorokhoz tar-

tozó listák halmazából. Részletes strukturáját a következőkben adjuk meg.

3.2. L lista, $k \geq 0$ egész szám, t a lista feje. Az L lista <u>k-adrendü</u> (jobb) <u>al-listájának</u> nevezzük azt az $L' = L \uparrow k$ listát, amelynek lista-feje t' (ha L' egyetlen atom, akkor t' ezt jelöli) és van olyan $a_0, a_1, \ldots, a_k, a_{k+1}$ cellasorozat (a_{k+1} atom is lehet), hogy $a_0 = t, a_{k+1} = t'$, minden i-re ($i = 0, 1, \ldots, k-1$) a_{i+1} a_i-nek jobb-komponense és a_{k+1} a_k-nak bal-komponense.

Tehát $L \uparrow 0$ a lista-fej bal-komponense, $L \uparrow 1$ a lista-fej jobb komponensének bal-komponense, és igy tovább. $L \uparrow k$ értelmezését kiterjeszthetjük olyan k számokra is, amelyekhez nem található a definicióban szereplő a_i sorozat (a jobb-indexek mentén haladva atomhoz jutunk $p \leq k$ lépésben): $L \uparrow k$ értéke legyen ilyenkor a 0 atom.

A <u>d</u> deszkriptorhoz a tezaurusz asszociációs listája rendelje az L listát. Az $L \uparrow k$ lista fejéből elérhető atomok halmazát <u>d k-adrendü értékkészletének</u> nevezzük és d_k-val jelöljük.

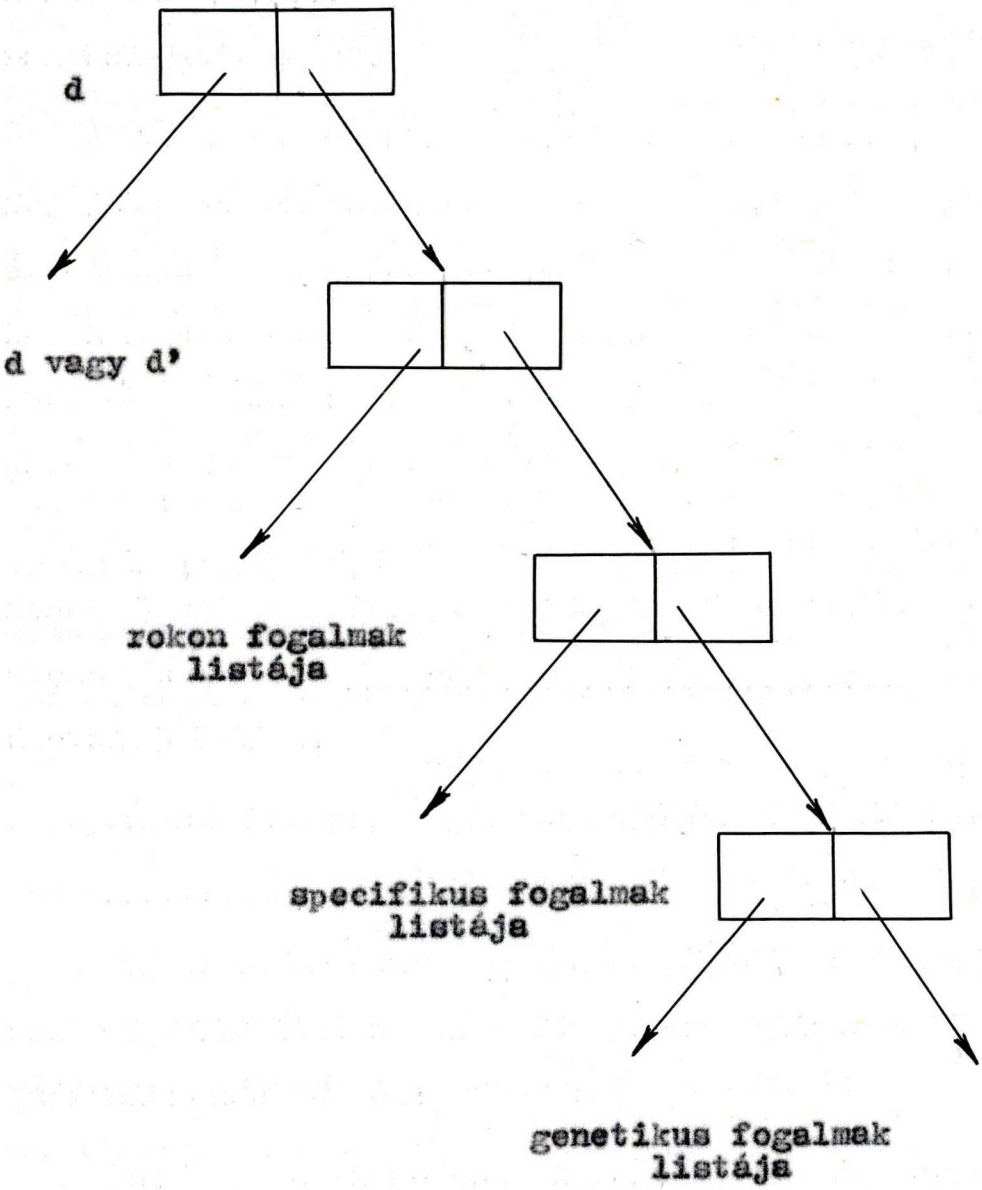
Általában az L lista fejéből elérhető atomok halmazát <u>L értékkészletének</u> nevezzük és L_H-val jelöljük.

A tezauruszban a <u>d</u> kulcsszóhoz, mint azonositóhoz az asszociációs listán keresztül rendelt lista a következő struktuárval rendelkezzék:

a./ Ha d nem-deszkriptor, akkor d_0 egyetlen deszkriptorból álljon, mégpedig a d-nek fogalmilag megfelelő d' kitüntetett deszkriptorból. Ha d deszkriptor, akkor

$d_0 = d$.

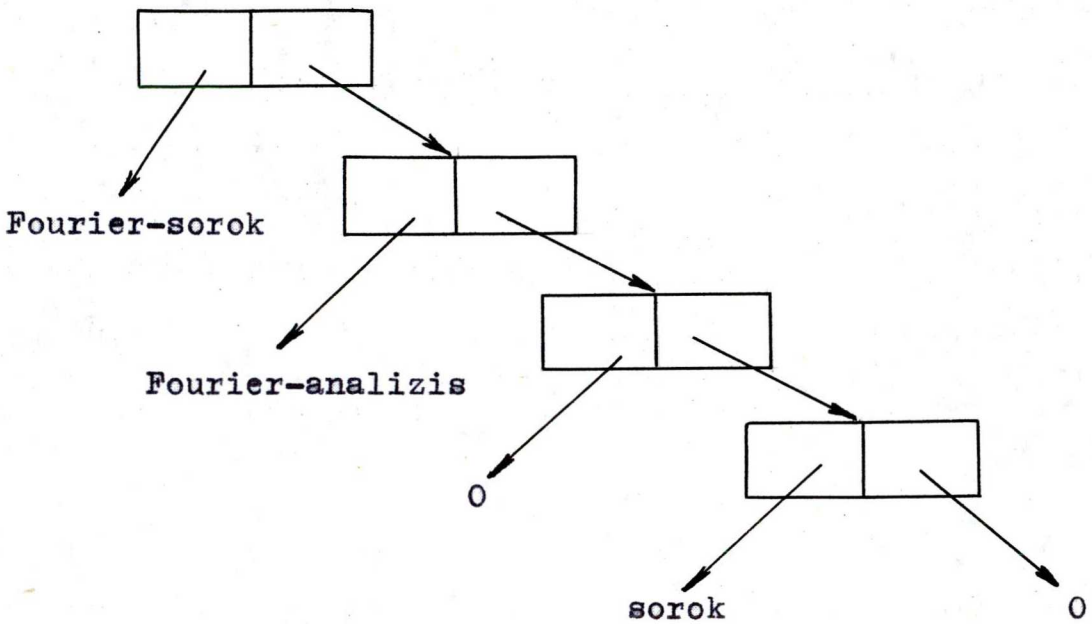
b./ d_1, d_2, d_3 rendre a rokon fogalmak, specifikus fogalmak, genetikus fogalmak deszkriptorainak halmazai legyenek. További fogalom-kategóriák felvételével a struktúrában $d_4, d_5, \ldots$ is kaphat szerepet.



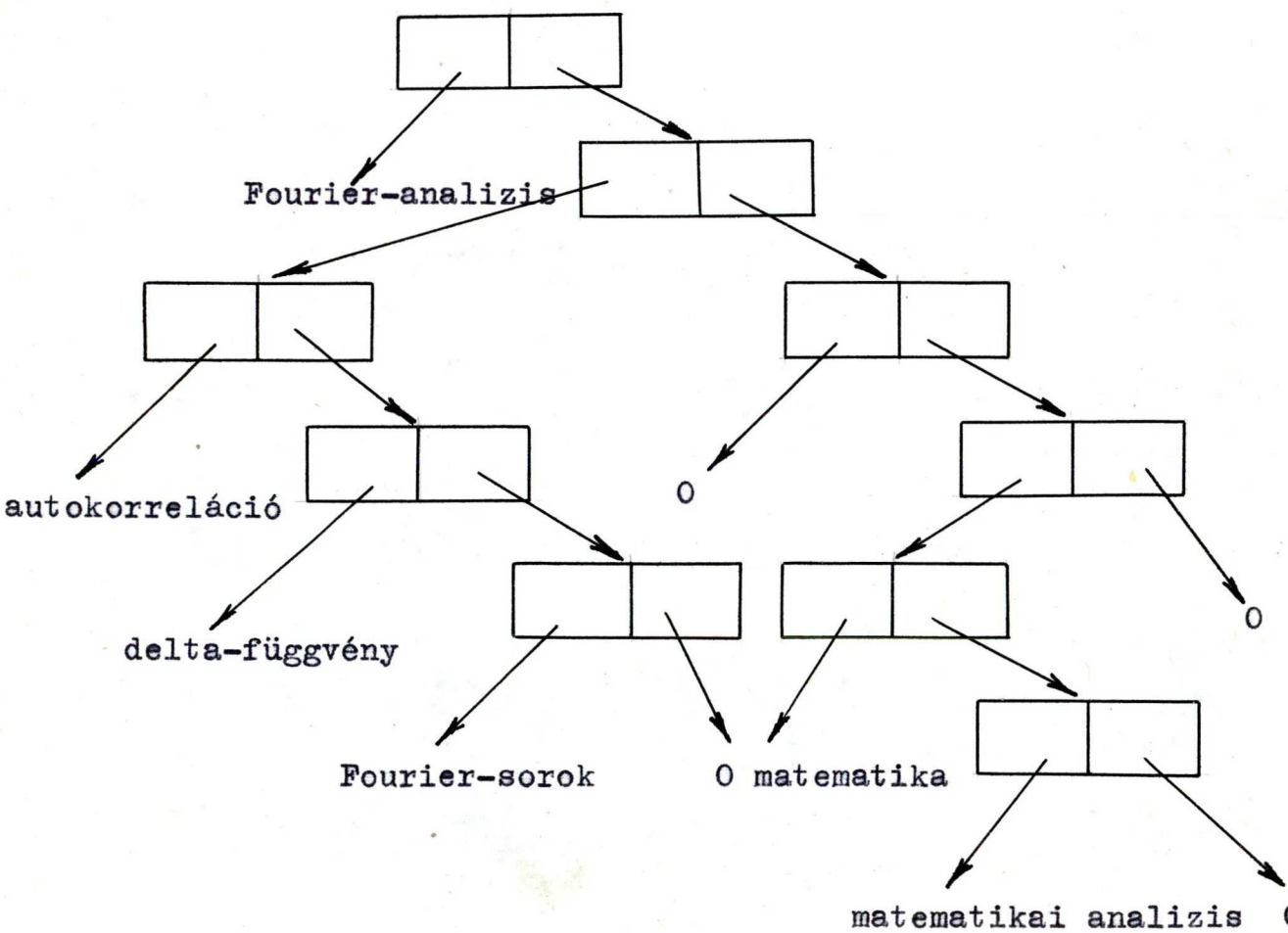
o

3.3. Példák

1./ Fourier-sorok:



2./ harmonikus analizis:



3./ Fourier-analizis: listája azonos a 2./ listával.

3.4. A tezaurusz számitógépi hordozója közvetlen elérésü háttérmemória lehet: mágneslemez, mágnesdob. A szekvenciális elérhetőség (pl. mágnesszalagos tárolás esetén) nem megfelelő még akkor sem, ha maguk a listák egy-egy rekordot képeznek: részben a deszkriptorok ismétlődései miatt (több tizszeresére növelné a tezaurusz terjedelmét), másrészt az asszociációs lista kifejezetten a direkt elérésü adatszervezés "katalógusának" szerepét tölti be.

A kombinált információvisszakereső rendszerre gondolva a tezaurusz strukturája a következő:

a./ Az asszociációs lista egy eleme

$$(d, L, I_1, I_2, \ldots, I_n)$$

alaku. d a kulcsszó szövege, L a kulcsszóhoz rendelt listából képzett rekord közvetlen elérési cime. $I_1, I_2, \ldots, I_n (n \geq 0$ azok a dokumentum-sorszámok, amelyekhez a d kulcsszót hozzárendeltük tartalmi feltárásuk, vagy a dokumentum-file-ba lépésükkor. Ilymódon az asszociációs lista betölti a kombinált információvisszakereső rendszer deszkriptor-file szerepét is.

b./ Minden egyes lista egy rekordot képez a tezaurusz adathordozóján . A cellák jobb-indexei mindig rekordon belüli valamilyen másik cellára vagy a 0 atomra utalnak. A bal-indexek vagy rekordon belüli cellára vagy az asszociációs lista egy elemére (azaz deszkriptorra) mutatnak. A cellák tehát nem szimmetrikusak, a jobb-indexek tárolására kevesebb bit is elegendő. Másrészt a cellák ilyen felépitése biztositja, hogy egy index csak akkor vezet ki a rekordból, ha atomra mutat, igy a tényleges lista-müveletekhez szükséges információk a

rekordon belül megtalálhatók. (Még az atomok összehasonlítására is elegendő: minden atom (deszkriptor) csak egyszer szerepel az asszociációs listában, ezért a rájuk mutató indexek egyenlősége az atomok egyenlőségével azonos.)

3.5. Az információvisszakereső rendszerben a (felhasználói) igény megfogalmazására (ld. 1.2.) tezaurusz birtokában a következő nyelv alkalmas:

3.5.1. A változók azonositóinak szerepét a kulcsszavak töltik be. A változó értéke az asszociációs listán keresztül a kulcsszóhoz rendelt lista.

3.5.2. Feltétlen és feltételes lista-kifejezéseket a változókból, lista-müveletekből (cons, hd, $t\ell$, set, sethd, settℓ, $\uparrow$,...), logikai értékü müveletekből (atom, eq,...) épithetünk fel. Minden lista-kifejezés eredményeképpen egyetlen lista alakul ki, ez a kifejezés értéke.

3.5.3. Az igény lehet (rekurziv definició):

a./ egy lista-kifejezés,

b./ igény negációja,

c./ igény kerek zárójelek között,

d./ két igény közé a konjunkció vagy diszjunkció jelét téve.

3.6. Szemantikailag az L lista az igényben az L érték-készletének L_H halmazát jelöli. Ha az igény egyetlen L lista-kifejezésből áll, akkor az igényt mindazok a dokumentumok kielégitik, amelyekhez hozzárendeltünk az L_H halmazból való deszkriptort.

Az L^1 és L^2 lista-kifejezés között a diszjunkció jele

az L_H^1 és L_H^2 halmazok egyesitését jeleneti.

Az $L^1 \wedge L^2$ igényt mindazok a dokumentumok kielégitik, amelyekhez L_H^1-ből is és L_H^2-ből is rendeltünk deszkriptort.

A $\neg L$ igényt azok a dokumentumok elégitik ki, amelyekhez L_H-ből való deszkriptor nincs rendelve.

Természetes módon definiáljuk a $\neg L^1 \wedge L^2$, $\neg L^1 \wedge \neg L^2$ $\neg L^1 \vee L^2$, $\neg L^1 \vee \neg L^2$ igények kielégithetőségét is. Ezeknél összetettebb igény-kifejezésekre a kielégithetőséget a konjunktiv vagy diszjunktiv normál-formára hozott alakon definiáljuk (a változók szerepét a lista-kifejezések illetve az általuk felépitett listák értékkészleteinek halmazai töltik be).

3.7. Az <u>információvisszakeresési eljárás</u> a következő lépésekben hajtható végre:

a./ Az igény-kifejezést konjunktiv vagy diszjunktiv normál-formára hozzuk (a benne szereplő lista-kifejezéseket azonositókkal helyettesitve).

b./ Az igényben szereplő kulcsszavakhoz az asszociációs lista segitségével kikeressük a tezaurusz megfelelő listáit.

c./ Az igényben szereplő lista-kifejezések értékét képviselő deszkriptor-halmazokat előállitjuk.

d./ A deszkriptor-halmazokban szereplő deszkriptorok (asszociációs listában tárolt) dokumentumsorszámait összegyüjtjük és növekvő nagyságrendben egy I sorozatot képezünk belőlük.

e./ Az I sorozat minden egyes elemének megfelelő dokumentum-rekord deszkriptorsorozatát összehasonlitjuk az igényt képviselő konjunktiv vagy diszjuntkiv normálformával. A normál-

forma azonositói helyett a nekik megfelelő deszkriptor-halmazokkal dolgozunk a 3.6. szabályai szerint.

Példa. Legyen a felhasználói igény-kifejezés a következő alaku:

cons(differenciálegyenletek↑0, differenciálegyenletek↑2)∧
cons(differenciálanalizátorok↑0, differenciálanalizátorok↑1)

Az igényben szereplő kulcsszavak mellett az Általános müszaki fogalmak tezauruszában a következőket találjuk (a lista-formában való megadás áttekinthetetlenebb, ezért felsorolás-formában adjuk meg):

differenciálegyenletek

specifikus fogalmak — lineáris differenciálegyenletek
parciális differenciálegyenletek

genetikus fogalmak — egyenletek

rokon fogalmak — Bessel függvények
differenciálhányadosok
differenciálok (matematika)
exponenciális függvények
Fourier-analizis
integrálegyenletek
kezdeti feltételek
Laplace-transzformáció
számitási eljárás
vektor analizis

differenciálanalizátorok→analóg számitógépek

analóg számitógépek

genetikus fogalmak — matematikai gépek
számitógépek

rokon fogalmak analizátorok (elektromos)
analóg-digitális átalakitók
analóg-digitális számitógépek
differenciálegyenletek
digitál-analóg konverterek
digitális számitógép
elektronikus számitógép
folytonosság
integrálók
logikai áramkörök

A két lista-kifejezést, az igényben való szereplésük sorrendjében jelöljük a-val és b-vel. A tezaurusz alapján a-nak és b-nek megfelelő listák fejéből elérhető atomok (deszkriptorok) halmazát jelölje A és B, amelyek a következők:

A = {differenciálegyenletek, lineáris differenciálegyenletek, parciális differenciálegyenletek}

B = {analóg számitógépek, Bessel-függvények, differenciálhányadosok, differenciálok (matematika), exponenciális függvények, Fourier-analizis, integrálegyenletek, kezdeti feltételek, Laplace-transzformáció, számitási eljárás, vektor analizis}

Tegyük fel, hogy a számitógépben tárolt tezaurusz aszszociációs listájában A és B elemei mellett az $i_1, i_2, \ldots, i_n$ indexeket találjuk meg. Ekkor a dokumentum-file $i_1, i_2, \ldots, i_n$ indexü rekordjaiban található deszkriptorsorozatot hasonlitjuk össze A és B elemeivel. Közülük egy rekord akkor és csak akkor felel meg az igénynek (amely végül is $A \wedge B$ alaku),

ha tartalmaz A-ból is és B-ből is deszkriptort.

3.8. A dokumentum-rekord információs rendszerbe lépésekor kulcsszavait az asszociációs lista alapján ellenőrizzük. Ha a d kulcsszó nem-deszkriptor, a tezauruszban hozzárendelt lista d_0 0-adrendü értékkészletének megfelelő d' deszkriptorral helyettesitjük. Ehhez le kell olvasni a listarekordok közül is némelyeket. Ha azt akarjuk, hogy elegendő legyen csupán az asszociációs lista alapján beépiteni a dokumentum-rekordot, a nem-deszkriptorokhoz rendelt asszociációs lista-elemet

$$(d, L, d')$$

formában kell tárolni: a dokumentum-sorszám sorozat helyébe (nem-deszkriptor esetében ez ugyis üres) az ekvivalens jelentésü deszkriptor kerül.

Nem elegendő azonban a dokumentum-rekord beépitéséhez a tezaurusz asszociációs listája, ha nemcsak a nem-deszkriptor kulcsszavakat helyettesitjük az ugyanazt a fogalmat jelölő deszkriptorral, hanem automatikusan a genetikus (esetleg rokon) fogalmak deszkriptorait is hozzá kivánjuk csatolni. Formálisan ez azt jelenti, hogy a d kulcsszó helyébe a d_0 és d_3 (esetleg d_1) deszkriptor-halmazok elemeit tesszük. Az ilymódon keletkező deszkriptorsorozat azonos tagjai közül természetesen csak egy kerül a dokumentum rekordba.

4. További lehetőségek

4.1. A működő információs rendszerek egy részében a dokumentum-rekordokhoz tartalmi kivonat és nem deszkriptorsorozat tartozik. A visszakeresési eljárás sebességének fokozása érdekében a dokumentum-rekord belépésekor a kivonatot szavakra tagolják és a továbbiakban, mint deszkriptorokkal dolgozik velük a rendszer. A kivonat tagolására két módszer kínálkozik:

a./ A rendszerbe építünk egy irreleváns szavakból álló szótárt (névelők, kötőszók, stb.). A kivonat minden olyan szava deszkriptor lesz, amely a szótárban nem szerepel. Hátránya, hogy a ragozott alakok külön-külön deszkriptort képviselnek, amit a felhasználói igények megfogalmazása küszöböl ki. (Csak a szótőt jelöli, amihez bármilyen ragot elfogad. Pl. asztal* jelöli az asztal, asztalt, asztalnak,... szavakat.) Másrészt a tartalomra nézve közömbös szavak is deszkriptorok lesznek, ami a terjedelmet növeli, egyuttal a visszakeresést lassitja.

b./ A rendszerbe a releváns szavak (kifejezések) szótárát épitjük és a tartalmi kivonat minden olyan kifejezése deszkriptor lesz, amely a szótárban (legalábbis tő-alakban) megtalálható. Tezaurusz birtokában ez az ut járható: a dokumentum-rekord belépésekor a tartalmi kivonat "értékelését" elvégezve, a visszakeresési eljárás alapját továbbra is a tezaurusz képviseli.

A tartalmi kivonathoz hasonló problémákat vet fel az igények természetes nyelven történő megfogalmazása, azzal a különbséggel, hogy bizonyos grammatikai információkra feltét-

lenül szükség van (és, vagy, nem szerepe, szavak sorrendje, stb.).

4.2. Az információs rendszer felhasználói igényeket értékelő része lényegében lista-kifejezéseket interpretáló program. A 2. és 3. pontban felsorolt lista-müveletek tovább bővithetők: függvények definiálhatók, rekurziv eljárások irhatók fel, stb. Lényeges szerephez juthatnak a standard módon beépitett makró-definiciók:bonyolult kifejezések az igények megfogalmazásakor csupán a makro nevére és a paraméterekre hivatkoznak. Pl. téma(d) jelentse a cons(d↑0, d↑1) kifejezést, ami a visszakeresésben a d deszkriptort és az összes rokon fogalmakat jelenti.

A makro-rendszer bővitései illetve a függvény-irásmód helyettesitése "szabadabb" formákkal, elvezethetnek a természetes nyelveken megfogalmazható igények közelébe.

Irodalomjegyzék

1. Általános müszaki fogalmak tezaurusza.
   Országos Müszaki Könyvtár és Dokumentációs Központ, 1970.

2. An Introduction to COMIT Programming. The Research Laboratory of Electronics and the Computation Center. The M.I.T. Press, 1962.

3. H.D.Avram, J.F.Knapp, L.J.Rather: The MARC II. Format. A communications Format for Bibliographic Data. Information Systems Office, Library of Congress. Washington, D.C. January, 1968.

4. N.G. de Bruijn: The mathematical lanuage AUTOMATH, its usaga, and some of its extensions. Proceedings of the Symposion on Automatic Demonstration.IRIA, Versailles, 1968.

5. R.G.Busacker, T.L.Saaty: Véges gráfok és hálózatok; bevezetés alkalmazásokkal. Müszaki Könyvkiadó, Budapest, 1969.

6. A.Church: The Calculi of Lambda-Conversion. Annals of Mathematics Studies № 6. Princeton University Press, 1941.

7. J.M.Foster: List Processing. Macdonald Computer Monographs, 1968.

8. R.R.Freeman, P.Atherton: File Organization and Search Strategy using the Universel Decimal Classification in Mechanized Reference Retrieval Systems. F.I.D./I.F.I.P. Conference on Mechanized Information Storage, Retrieval and Dissemination. Rome, June, 1967.

9. J.Friant: Les Languages "Contex-Sensitive". Ann. Inst. Henri Poincaré, Vol. III. № 1, 1967. p. 35-120.

10. M.A.Gavrilov, A.D.Zakravskii: LYaPAS, A Programming Language for Logic and Coding Algorithms. Academic Press, 1969.

11. В.М. Глушков: Синтез цифровых автоматов. Государственное Издательство Физико-математической Литературы. Москва, 1962.

12. T.M.Hibbard: Some Combinatorial Properties of Certain Trees with Applications to Searching and Sorting. Journal of the ACM. January 1962.

13. Horváth Tibor: A könyvtárak müszaki fejlesztése. Budapest-Veszprém, 1969. OSzK, Könyvtártudományi és Módszertani Központ.

14. IBIS; System wyszukiwania informacji. Instytut Maszyn Matematycznych. 1968.

15. INBI; System przetwarzania na EMC informaji bibliograficznych Instytut Maszyn Matematyeznych. 1968.

16. INFORGA Szimpozium előadások tézisei. Moszkva, 1965. Országos Müszaki Könyvtár és Dokumentációs Központ, 1965.

17. L.Kalmár: Les Calculatriees automatiques comme structures algériques. Gauthier-Villars-Éditeur. 1963.

18. Kalmár László: Matematikai és nyelvi strukturák. Általános Nyelvészeti Tanulmányok II. Akadémiai Kiadó, Budapest, 1964

19. C.A.Kapps: SPRINT - A programming Language with General Structure. The Moor School of Electrical Engeneering. Report N° 71-18. 1970.

20. Makay Árpád: Tanulmány a kisszámitógépek könyvtári és dokumentációs alkalmazásait illetően. JATE Kibernetikai Laboratóriuma, 1971.Kézirat.

21. Nicht-Numerische Informationverarbeitung. R.Gunzenhäser. Springer Verlag, Wien-New York. 1968.

22. J.O'Connor: Mechanised Indexing Methods and Their Testing. Journal of the ACM. V.11.N° 4. October 1964.

23. D.D.Prentice: The Combined File Search System. IBM Corporation, 1964.

24. Scientific, Technical and Economic Information in Poland. Central Institute for scientific, Technical and Economic Information. Warsawa, 1963.

25. Searching Normal Text for Information Retrieval. IBM Data Processing Application, E20-0335-0. 1969.

26. SESAM. Kézirat.

27. C.Weissmann: LISP 1.5. Primer. Dickenson Series in Computer and Information Science.

Tartalomjegyzék