is megérdemli, hogy a szó különös értéket jelző értelmében mühelynek nevezzék. A kötetek - jól sikerült -"esztétikai gondozása" (boritólap,belső illusztrációk) Gonda Zoltánra hárult.

FUTALA Tibor

K.Leon MONTGOMERY: Document Retrieval Systems. Factors Affecting Search Time. Marcel Dekker, Inc. New York. 1975. VIII. 144 p.

DOKUMENTUMVISSZAKERESŐ RENDSZEREK. A KERESÉSI IDŐT BEFOLYÁSOLÓ TÉNYEZŐK

A visszakereső rendszerek hatékonyságának kérdése napjainkban élő probléma. A Pittsburgh-i Egyetemen (Pennsylvania) kisérletsorozatot folytattak, amelynek eredményeit a jelen - könyv formában publikált - kutatási jelentésben teszik közzé.

A géppel olvasható adattárakban való kereséssel kapcsolatos problémák a hatvanas évek legelején vetődtek fel. Ezek a gépek azonban még meglehetősen kis központi memóriával rendelkeztek, a háttértár gyakran néhány mágnesszalagot vagy egy mágneslemezt jelentett csupán. 1965-ben megjelentek a harmadik generációs gépek, amelyek lehetőségeket nyitottak meg az információ visszakeresés terén is. Három igen lényeges kérdésben hoztak ujat, kettőt ebből a hardware , egyet software területén: Hardware tekintetében gyorsabb processzort alkalmaztak és jelentősen megnövelték a belső és külső tárak kapacitását. A harmadik jelentős ujitást az operációs rendszeren hajtották végre.

Már jóideje köztudott, hogy a visszakeresési időt nagy mértékben befolyásolja az adattár nagysága és szervezése. A kisérlet végrehajtói vállalkoztak arra, hogy bemérik a keresési időt befolyásoló összes egyéb paramétert, a költségtényezők függvényében, igy megvizsgálták a file szervezési módszereket, a kereső programokat, és a rendelkezésre álló számitógép rendszert. A géppel olvasható adattárak használata csak akkor gazdaságos, ha lehetőleg minél kisebb költséggel minél több kérdést tudnak megválaszolni, minél több felhasználót tudnak kiszolgálni.

Vizsgálataikat a file szervezés tanulmányozásával kezdték. A file szervezésnek két alap formája ismeretes: a lineáris és az invertált. Lineáris file szervezés esetén a dokumentumok rekordjait az azonositó számok sorrendjében tárolják. Minden dokumentumot jellemeznek index kifejezésekkel, és ezeket a dokumentumhoz rendelik. Ez a következőképpen érzékeltethető:

Ha az

1.sz. dokumentumot A, B, F és G index kifejezések jellemeznek és a
2.sz. dokumentumot B, G, és H az adattár szerkezete a következő:
1.sz. dokumentum A, B, F, G index fogalmak
2.sz. dokumentum B, G, H index fogalmak.

Invertált file szervezésekor az index fogalmakat rendezik sorba és minden indexfogalom után felsorolják azokat a dokumentumokat, amelyekben a kérdéses indexfogalom előfordul. Az előbbi példánál maradva, az invertált file a következőképpen épül fel:

A indexfogalom	1.sz. Dokumentum
B indexfogalom	1.sz. Dokumentum, 2.sz. Dokumentum
F indexfogalom	1.sz. Dokumentum
G indexfogalom	1.sz. Dokumentum, 2.sz. Dokumentum
H indexfogalom	2.sz. Dokumentum

A lineáris file szervezés lényegesen könnyebben megvalósitható, igen egyszerü az ujabb adatok beépitése (update-olás), könnyen követhető a keresés logikája, és nem utolsó sorban batch üzemmódban igen gazdaságos. Valószinüleg ez az oka annak, hogy ez a file szervezési módszer terjedt el szélesebb körben. (NASA Technology Transfer Project, Chemical Abstracts Serv.)

Minél több kérdésre keres egyidőben azonos file-on, annál kisebb az egy kérdésre eső költség.

Megfigyelték, hogy azonos nagyságu file esetén a keresési időt nagyon befolyásolja a keresési profilok száma és a két emlitett file szervezés módszer, ezért a kisérletek során kidolgoztak egy olyan speciális kereső programot, amellyel a lineáris és invertált szervezésü file-on egyaránt lehet visszakeresni. A programot a Pittsburgh-i Egyetem IBM 360/50-es számitógépére irták.

A munkálatokat a kérdés szakirodalmi felderitésével kezdték, az 1960-as évektől szemleszerüen feldolgozott anyagot a 2.sz. fejezetben ismertetik.

Az irodalomban tárgyalt kérdéseket végül három csoportba rendezték: vizsgálták a keresési időt:

- az egyidejüleg futó kérdések számának függvényében,
- az adatbázis méretének függvényében
- és az alkalmazott file szervezés tekintetében.

Ugy szervezték tehát a kisérleteket, hogy a fenti kérdésekre választ kapjanak, ezért a következő feltételeket kellett biztositani a kisérletek lefolytatásához:

1. az adatbázis létező, széles körben ismert és használt legyen,
2. aktiv, tényleges, valós kérdésekkel fusson a program,
3. az adatbázis lineáris és invertált file szervezést is reprezentáljon vagy ilyenné átalakitható legyen,
4. és végül, hogy az futtatható legyen az Egyetem adott gépein.

A kisérletek szempontjából két gépi adattár jöhetett szóba a National Aeronautics and Space Administation (NASA) Technology Transfer Project szalagjai vagy a National Science Foundation (NSF) támogatásával a Chemical

Information Center által kiadott Chemical Abstracts Services Condensates adattára. Némi vizsgálódás után végül is a CAS szalagot választották, mivel az Egyetemen ezt az adattárat az IBM 360/50-es gépen futtatják, amely nagyobb kapacitásu, mint a NASA adattárral dolgozó másik számitógép. Az IBM gép lehetőséget nyujtott a multiprogramozási lehetőségek tanulmányozására is.

A kisérlet leglényegesebb alkotó elemei: az adatbázis, a kereső profilok, a kereső program és a számitógépes rendszer. Az általuk használt adatbázis, amely 74. volume 2. issu szalagváltozata öt adatcsoportot tartalmaz:

1. A dokumentum, a kötetszám és füzetszám,
2. A cim
3. A szerző
4. A szerzővel kapcsolatos adatok
5. Az index fogalmak.

Ezek közül csak három adatcsoportot használtak: 1. A dokumentum (kötet és füzetszám); 2. A cimet; 3. Az index fogalmakat.

A cim és az indexfogalmak alapján végzik a keresést, és határozzák meg a kötet és füzetszámot. Általában 10-50 fogalom állt rendelkezésre a dokumentum keresésére és kb. 20 terminussal kerestek minden dokumentumra. Az eredeti szalagon (adatbázison) az adatok lineárisan vannak szervezve. A kérdéseket a Chemical Information Center Project kb. 400 kérdéséből 180-at választottak ki, majd végül 128 kérdéssel végezték a kisérletet, Boole logikai elemek (és, vagy, és nem) alkalmazásával. Az IBM 360/50 operációs rendszere lineárisan szervezett file-ra iródott, módosithatták volna ezt is, de végül is egy uj, az invertált szervezésü file-on való keresést lehetővé tevő programot irtak.

A kisérlet során két módszert alkalmaztak: először megvizsgálták a keresési idő és a keresett dokumentumok viszonyát, majd a kérdések számának függvényében vizsgálták a keresési időt. Természetesen mindkét variációt: lineáris és invertált szervezésü file-on is kipróbálták, az eredményeket összevetették. A dokumentumok számának és a keresési időnek vizsgálatakor 1, 256, 512, 1024, 2048 és 4096 dokumentumot kerestek vissza, lineáris és invertált file-on egyaránt. A kérdések számát a kisérlet második felében pedig a következőképpen változtatták: 1, 16, 32, 64, 96 és 128 kérdést tettek fel. Amikor lehetőség nyilott, háromszor megismételték a futtatást, hogy igy átlag értékekhez jussanak.

A kisérlethez un. bimoduláris programot dolgoztak ki, amely egyaránt használható invertált és lineárisan szervezett adattárak esetében. Olyan programot kellett irni tehát, amelyben lehetőleg azonos szubrutinokat alkalmaznak és csak, ahol feltétlenül szükséges, tér el egymástól a visszakereső program, mert egyébként a különböző program lépésekből, megoldásokból is származnak időeltolódások.

Részletesen ismertetik, a számitógép tipusát, lényeges és jellemző adatait, a configurációkat, a program minden jellemzőjét, a használt kifejezéseket, az index fogalmakat és minden lényeges adatot.

A kisérleti futtatások során a következő adatokat regisztrálták automatikusan:

- a lineárisan szervezett file-on vagy az invertált file-on kerestek,
- az átnézett dokumentumok száma (1-4096)
- hány kérdést tettek fel (1-128)
- a futtatás pontos ideje és napja
- hány kereső profilt használtak
- a futtatás száma
- a központi memória időigénye
- input és output idők
- teljes számitógépes idő
- a futtatás költsége
- az output sorok száma
- a keresésre előkészitett file invertálás ideje
- a 4096 dokumentum szalagról lemezre másolásának ideje
- visszakeresési idő
- a sikeresen visszakeresett dokumentumok száma
- és hogy mennyi idő kell az output perifériák előkészitésére.

Ezeket az adatokat egy 80 oszlopos lyukkártyára vitte fel a gép, minden egyes futtatásról készitettek ilyen kártyát, összesen 321-et, a kisérlet hat hete alatt.

A számadatokból szerkesztett ábrák és diagramok alapján a következő megállapitásokat szürték le: a visszakeresési idő egyenes arányban változik a dokumentumok számával, kezdetben. Kisebb adatbázisnál lineárisan szervezett adattárban érhető el a legrövidebb idő alatt a kérdéses dokumentum. Nagyobb mennyiségü adat kezelésére az invertált szervezés ajánlott. Arra, hogy hol van ez a metszéspont, amelyből megállapitható, hogy hol gazdaságos egyik és honnan már a másik szervezési módszer - az egyidőben futó kereső profilok száma van nagy hatással. Minél nagyobb számu kérdés fut egyidőben, annál kisebb adatbázis esetében az invertált file szervezés eredményesebb. Ha az egyidőben futtatott kereső profilok száma növekszik, akkor már kisebb adatbázis esetében is előnyösebb az invertált file alkalmazása.

A kisérlet során megvizsgálták a dokumentumok számának és a kérdések számának viszonyát is. Adott számu dokumentumbázis esetén az egyidejüleg futó kérdések számának növekedésekor elérkezünk egy olyan kérdésszámhoz, amelynél nagyobb számu kérdés esetén már az invertált szervezésü adattárnál k kisebb az egy-egy kérdésre jutó idő.

A vizsgálatok egyik megállapitása, hogy a visszakeresést célszerü batch üzemmódban végezni, igy a legnagyobb a hatékonyság. Ugy találták, hogy 64 kereső profil egyidejü használatakor kapták a legrövidebb egységnyi keresési időt, ha ennél többet kerestek egyidejüleg az egységnyi keresési idő növekedett.

Érdekes megállapitás az is, hogy célszerü az invertált szervezésü filelal dolgozni, ha 32-nél több kérdés fut egyidőben, és ha az adott adatbázis nagysága meghaladja az 512 dokumentumot.

A kisérletek során számos értékes eredmény és megállapitás született, de egyben ujabb kérdések is felvetődtek. Feltétlenül meg kell vizsgálni a közeljövőben a fileszervezés egyéb más megoldásait, lehetőségeit. A munkálatok fő eredményének a bimoduláris STRESS program kialakitása tekinthető.

BOBOKNÉ BELÁNYI Beáta