METODA PRENOSA VELIKE MNOŽICE VEKTORJEV VREMENA NA HITRE MEDIJE RAČUNALNIKOV

A METHOD FOR TRANSFER OF LARGE WEATHER VECTORS' POPULATION TO HIGH SPEED INPUT MEDIA OF COMPUTERS

551.582.2 : 681.327.4/.6

A. HOČEVAR, M. JURGELE * in Z. PETKOVŠEK

Univerza v Ljubljani in
* Iskra Standard, Ljubljana

SUMMARY:

In this paper a method is given for transfer of large weather vectors' population (order of 10 millions) from low speed input media (cards) to high speed ones (magnetic tapes).

The program is written in FORTRAN IV language and transfers for instance, meteorologic data from cards to magnetic tapes in two-dimensional arrays, with dimensions 31 and 23. In such an array where data are packed if needed, to take as little space as possible, data are being gathered together for a whole month. If the month has less than 31 days, the array is completed with symbolic values.

The right time sequence of data is tested in such a way that the first and the last day of the month have to be on the right place of the time scale. The wrong time sequence and problems, connected with too many and not enough cards, are solved by the computer. Temporary arrays of same dimen-

sions are formed and completed by symbolic values. Detailed informations about them are printed on the printer. Thus mistakes can be eliminated later. The catching of the right time sequence again causes the transfer to follow the original schema without remarks (Fig. 1).

The problem with space started at the cards already. They could not be punched according to the numerical code of the computer but can be punched after additional code. Large numbers written in more than one column or other informations can be written in less columns with the help of rows, which aren't used in the numerical code of the computer. A suitable program for reading such informations was written and basic informations about it are given.

The final result of this work - a rational record of data on high speed input media-enables a quicker and cheaper further research on computers.

IZVLEČEK

Shema prenosa velike množice komponent vektorjev vremena (nekaj milijonov), ki je podana v tem delu, je zgrajena tako, da teče prenos brez zastojev ter da stroj upošteva in tudi sam rešuje možne nepravilnosti in jih signalizira. Vektorji so razporejeni v paketih po 31 krat n podatkov ter pakirani tako, da je prostor na hitrih medijih čim bolje izkoriščen. Poseben sistem omogoča čitanje kartic, ki so luknjane po posebnem kodu. Vse to omogoča, da je nadaljna obdelava podatkov z računalniki hitrejša in cenejša.

UVOD

Znanosti, ki se pri svojih raziskavah opirajo na velike množice podatkov, imajo dandanes vse večje možnosti za svoj razvoj. Med te vede spada tudi meteorologija. Medtem, ko so bile raziskave velikih množic podatkov še v pretekli dobi precej omejene zaradi počasnosti klasičnega načina dela, lahko podatke sedaj vsestransko obdelujemo z računalniki, seveda le, če so pripravljeni tako, da jih stroj lahko tudi prečita.

Pri velikem številu podatkov - velikostnega reda deset milijonov, ki jih v meteorologiji ne redko kompleksno obdelujemo, pa možnost čitanja sama ni dovolj. Pojavlja se problem, kako hitro stroj podatke lahko čita in kako lahko z njimi ponovno in ponovno manipuliramo. Če so podatki na luknjanih karticah (ali na perforiranem traku), jih stroj sicer lahko čita, vendar gre tako čitanje sorazmerno počasi in je manipulacija s stotisoči kartic zelo neprikladna.

Čitanje 20.000 podatkov na minuto, kar je za klasično pojmovanje mnogo, je za velike množice podatkov še vedno prepočasno. Hitri mediji, kot sta magnetni trak ali disk, omogočata ob primernem sistemu stokrat hitrejši prenos. Zato je potrebno prenesti podatke s kartic ali trakov na te hitrejše vhodno - izhodne medije; z njimi je delo mnogo hitrejše, bolj prikladno in cenejše.

Ta prenos je v načelu sicer preprost. Če pa upoštevamo, da je možno:

- da podatki - kartice - niso v pravilnem časovnem zaporedju in jih lahko nekaj manjka ali jih je celo preveč,
- da hočemo imeti zapis na hitrem mediju zelo racionalen in tak, da bomo podatke lahko hitro urejali in kombinirali, ter
- da podatki na počasnih vhodnih medijih zaradi stiske s prostorom niso luknjani v direktno čitljivem kodu,

pa postane prenos podatkov s počasnih vhodnih medijev na hitre manj preprost problem in njegova rešitev je pogoj za uspešno nadaljnje delo. Prav reševanju tega problema pa je posvečeno to delo.

OSNOVNI PROGRAM

Osnovno vodilo glavnega programa, kadar gre za delo z veliko množico podatkov je to, da med delom ne sme priti do zastoja. Kljub veliki hitrosti dela lahko trajajo operacije z desetmilijonskimi podatki več ur. Zato morajo biti predvidene vse možne neregularnosti, ki jih mora računalnik sam sproti rešiti in nadaljevati z delom. Pri tem se mora držati osnovne sheme, obenem pa preiti neregularnosti, jih vključiti v osnovno shemo in obenem nanje tudi opozarjati. Le tako so možne kasnejše hitre in racionalne korekcije in ureditev celotnega materiala.

Drugo vodilo je obseg hitrega spomina uporabljanega računalnika. Pri delu z veliko množico podatkov je ta spomin sorazmerno majhen ter je nujno potrebno delo v ciklusih, kot je bilo to pri nas tudi v zvezi z reševanjem meteoroloških problemov že obravnavano za podatke s sinoptičnih postaj /1/. Klimatološke postaje dajejo dnevno manj, to je, okrog 70 komponent vektorja vremena, velikost spomina večjih sodobnih računalnikov, ki sme biti največ do tretjine zavzet s podatki, pa je okrog 30.000 besed. Glede na to je za kompleksne klimatološke obdelave, vključno prostorsko razporeditev, primeren ciklus oziroma "paket" podatkov za en mesec.

Vzemimo, da imamo na počasnih vhodnih medijih (n.pr. na karticah) vektorje vremena $a = a(x_1, x_2, \ldots x_n)$, ki so časovno in krajevno definirani. Te želimo spraviti na hiter vhodni medij v paketih dvodimenzionalnega polja, katerega ena dimenzija bo velikosti 31, kolikor je največ dni v mesecu, druga pa velikosti n, kolikor je komponent tega vektorja, to je, kolikor imamo

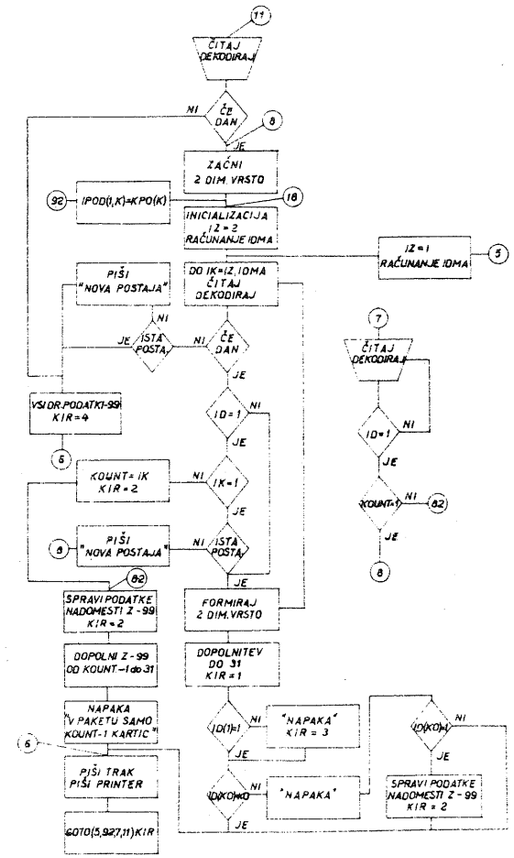
podatkov na dan. V vsakem takem paketu bo torej za en mesec podatkov, od katerih morata biti vektorja vremena vsaj za prvi in za zadnji dan na pravem mestu časovne premice. Podatki morajo biti za vse dni v mesecu. Za mesece, ki imajo manj kot 31 dni, se paket dopolni s simboličnimi vrednostmi. V primeru, da so te minimalne zahteve izpolnjene, bo tekel prenos gladko oziroma po osnovni shemi, sicer pa občasno po variantni. Šele ko bo prenos končan, je smiselno pričeti s celotno kontrolo datuma in vektorjev podatkov /1,2/ in z njihovim eventuelnim kompletiranjem /3/ kot naslednjo stopnjo obdelave.

Vektorji vremena morajo biti zapisani na hitrem mediju po enotni shemi, možno pa je, da podatki, ki jih stroj čita, niso nizani v pravilnem časovnem zaporedju. Zato vključimo v program navodilo, kako naj računalnik izpiše odstop dejanske razporeditve podatkov od osnovne sheme. Paketi, ki so vedno dimenzij 31 krat n, so le pomožni, kadar so podatki neurejeni. Takih podatkov je praviloma malo, uredimo pa jih kasneje. Čim pride pri čitanju stroj spet na urejen material, mora iti prenos po osnovni shemi v redu naprej.

Shema poteka operacij z upoštevanjem zahtev v prejšnjem odstavku in možnosti za konkretno nalogo prenosa klimatskih podatkov z luknjanih kartic (po ena na dan za vsako postajo) na magnetni trak, je podana na sliki 1. Upoštevana je še informacija, da za mesec, za katerega vsi podatki manjkajo, ni kartic za vsak dan, ampak je samo ena za ves mesec, rubrika za dan na njej pa je prazna /4/.

Potek operacij je naslednji: Stroj čita prvo kartico. Če ni luknjana v direktno čitljivem kodu, jo najprej po podprogramu analizira, pretvori in nato posreduje v glavni program. Splošna rešitev tega problema je posebej podana v tretjem poglavju tega dela. Če je rubrika za dan prazna, napravi stroj paket 31 krat n simboličnih komponent ter začne čitati naslednjo kartico, kot da bi prejšnje sploh ne bilo. Če pa podatek za dan obstaja, formira iz podatkov te kartice prvo vrsto paketa, izračuna koliko dni ima tisti mesec ter toliko kartic tudi takoj prebere. Čitanje prekine, če naleti na datum 1, a o tem pozneje. Iz prebranih podatkov formira dvodimenzionalno polje ter ga dopolni do 31 krat n s simboličnimi podatki, če je dni v mesecu manj kot 31. Tako formirano polje v spominu prekontrolira glede na prvi in zadnji vektor in izpiše eventualne nepravilnosti oziroma odstope od osnovne sheme (n.pr. "druga postaja", "v paketu je samo k-3 kartic" itd.). Nakar šele zapiše celo dvodimenzionalno zbirko na trak v enem paketu in kot en "record".

Stroj nato izračuna koliko dni ima naslednji mesec. Če je vse v redu, prečita toliko kartic, formira zbirko, jo po potrebi izpolni in spet zapiše na magnetni trak. Če prva brana kartica ni tudi prvi dan v mesecu in je prvi dan šele na kaki drugi kartici, recimo na deseti, stroj iz prvih devetih prebranih kartic napravi pomožen paket in opozori na napake, s podatki desete kartice - s prvim dnevom v mesecu pa začne nov paket, ki je dokončen, če je najprej vse v skladu z zahtevami programa in spet začasen, če ni.



Slika 1 Diagram poteka operacij

Fig. 1 Flow chart

Primer, da kartica prvega dne manjka, je rešen tako, da stroj po kontroli dvodimenzionalnega polja, za katerega je ugotovil, da ne vsebuje prvega dne, ob čitanju kartic vsako testira, če je na njej prvi dan. Čim pride do nje, napravi iz že prebranih kartic pomožen paket, s kartico prvega dne pa začne novo dvodimenzionalno polje oziroma nov paket. Vsako spremembo lokacije (številko postaje) nam da stroj po testiranju vsakega prvega dne v mesecu, ki jo izpiše na printerju skupno z vektorjem vremena za ta dan, kar je potrebno zaradi kontrole poteka operacij, ki kljub veliki hitrosti računalnika trajajo več ur.

Da bi zaščitili že prebrane in na magnetni trak prenešene podatke pred izgubo v primeru okvare računalnika ali izpada električne energije i.p., so le-ti združeni v posamezne enote, ki so med seboj ločene s posebno znamko - file mark. Pri morebitni okvari zato ni potrebno ponavljati čitanja vseh podatkov, ampak se z vključitvijo stikala na konzoli računalnika in z dodatno kartico /5/ preskoči vnaprej znano število enot zapisa na traku, saj je le to po zapisu vsakega file tiskano na linijskem printerju, nakar se prenos nadaljuje.

RACIONALIZACIJA PROSTORA

Magnetni trak ali disk sta hitra magnetna medija, ki omogočata sama po sebi največjo koncentracijo podatkov, vendar pa lahko z uporabo dodatnih trikov in sistemov zapišemo na tak medij deset in več krat toliko podatkov kot brez njih. Pri veliki množici podatkov je to pomembno, saj lahko imamo na enem traku vse tisto, za kar bi jih sicer potrebovali deset. To pa je že zaradi prostora in prenosa, predvsem pa zaradi hitrejše ročne manipulacije velika prednost, ki je pri veliki množici podatkov ne smemo opustiti. Eden izmed tovrstnih posegov je pakiranje podatkov, drugi pa uporaba posebnih sistemov pisanja.

Vzemimo kot preprost primer, da je največje število decimalnega sistema, ki ga lahko zapišemo v prostor za besedo, šestmestno število

9 9 9 9 9 9

Večina komponent vektorja vremena oziroma meteoroloških podatkov pa je eno ali dvomestnih (n.pr. oblačnost $x_7 = 4$, stanje tal $x_8 = 1$, smer vetra $x_9 = 27$ itd.), ki so v prostore za besede napisani takole:

b b b b b 4 b b b b b 1 b b b b 2 7

kjer pomeni b prazen prostor (blanck). Vse te podatke lahko spravimo v prostor za eno samo besedo na primer takole:

b 4 b 1 2 7

in nova komponenta y_k ima vrednost 40127; z njo se je potrebni prostor zmanjšal za faktor 3.

Postopek za tako pakiranje je zelo preprost in ga napišemo simbolično;

a za naš primer:

$$y_k = 10000\ x_7 + 100\ x_8 + x_9$$

kjer so x_7, x_8, x_9 pozitivna cela in največ dvomestna števila (integerji > 0).

Če tako pakirane komponente zapišemo na hitrejše medije, je teh torej lahko trikrat manj; v isti obliki pa gredo seveda ob čitanju nazaj v hitri spomin računalnika. Da dobimo spet prvotne podatke x_7, x_8, x_9, ki so potrebni za nadaljnje delo, pa ne moremo preprosto po eksplicitnem zapisu enega iz - med sestavin izločiti, ker vsebuje gornja enačba sedaj tri neznanke in ni rešljiva. Tu uporabimo lastnost računalnikov, da pri prenosu definiranega integerja iz operativne enote v hitri spomin, vrednost, ki je za decimalno vejico, ignorira. Glede na to dobimo iz pakirane vrednosti y_k razpakirane komponente po enačbah, ki jih uporabimo po naslednjem vrstnem redu:

$$x_7 = y_k / 10000$$
$$x_8 = (y_k - 10000\ x_7) / 100$$
$$x_9 = y_k - 10000\ x_7 - 100\ x_8$$

Negativnih vrednosti na ta način ne smemo pakirati, ker je pri razpakiranju predznak komponent nedoločljiv. Pomagamo si lahko tako, da vsem vrednostim take komponente vektorja prištejemo določeno pozitivno konstanto, ki je zagotovo večja od katerekoli negativne vrednosti te komponente. S tem dobimo same pozitivne vrednosti in lahko dalje ravnamo po gornjem načinu. Pri končni določitvi posameznih vrednosti konstanto seveda spet odštejemo. V izjemnih primerih je možno pakirati celo po več besed v eno samo in tako ustrezno zmanjšati zasedbo hitrega medija. Pri majhnem številu podatkov to ni smiselno, je pa zato tem bolj, če gre število podatkov v desetine milijonov oziroma tedaj, ko podatki presežejo obseg fizične enote medija.

Posamezni "record" - zapis na traku - sprejme v navadni obliki le so - razmeroma majhno število karakterjev - znakov (med 120 in 160), za njim pa ostane vedno 18 mm traku praznega, tako da je več traku praznega kot polnega. Zato je potrebno pri veliki množici podatkov uporabiti sistem, ki bo omogočal zapis celega paketa v en "record", s čemer je torej lukenj tridesetkrat manj. Pri našem delu na elektronskem računalniku CDC 3300 smo uporabili sistem BUFFER /5/ in s tem reducirali dolžino magnetnih trakov za faktor 4.

VREDNOTENJE DIREKTNO NEČITLJIVIH INFORMACIJ

Zaradi racionalizacije prostora, ki je pogosto potrebna tudi na počasnih medijih - karticah, so kartice včasih luknjane tako (na pr. z dodatnimi luknjicami na posameznih pozicijah), da kombinacija luknjic ne ustreza nobenemu izmed numeričnih karakterjev računalnikovega koda. Taka dodatna kodiranc sporočila morajo biti, seveda, tako izbrana, da vrednosti enoumno določajo.

Obstajajo lahko za vsako kolono ali grupo kolon na kartici. Številke v posameznih kolonah ali grupah kolon so lahko indentifikacijske količine (n.pr. dan, mesec, postaja) ali različne vrednosti (n.pr. pritisk, relativna vlaga na kartici Klima 1, oblačnost in vremenski pojavi na kartici Klima 2). Kodirni sistem je v našem primeru tako izbran, da pokličemo na pomoč enajsto in dvanajsto horizontalno vrstico. Vse vrednosti na karticah Klima 1 in Klima 2 /6/ so numerične in so označene s kombinacijo števil od 0 do 9. Zaradi stiske razpoložljivega prostora se vrednosti, ki bi normalno zavzele več vertikalnih kolon, luknjajo na manj mestih, zato pa s pripadajočimi simboličnimi označbami X, Y v vrstah 11 in 12. N.pr. vrednost 12,5, ki bi pri normalnem načinu zapisa zavzela štiri vertikalne kolone, se lahko zapiše v skrajšanem zapisu na dveh kolonah takole $\overset{x}{25}$. Čitalec kartic v sklopu računalnika prebira kolono za kolono in vsako kombinacijo luknjanih mest prevede v svoj lastni računalniški kod na šestih binarnih mestih. Ta zapis ima svoje ime; Internal BCD zapis. Te prebrane veličine - karakterji so lahko v območju od $00_8 - 77_8$ in zavzemajo vse možne kombinacije, ki jih lahko zapišemo na šestih binarnih mestih. Številke zavzemajo spodnja mesta od $00_8 - 11_8$ (številke 0 do 9), ostali karakterji pa mesta od 11_8 navzgor. Vse vrednosti lahko razdelimo v tri grupe:

1. V tej grupi so vrednosti, ki so manjše ali enake 11_8. To so čista števila in tiste, ki so enake 60_8, to je neluknjanim mestom v koloni.
2. V tej grupi so vrednosti, ki so manjše ali enake 32_8, a večje ali enake 21_8. To so kombinacije luknjanih mest v horizontalnih vrsticah 0+12, 1+12, 2+12, 3+12, 4+12, 5+12, 6+12, 7+12, 8+12, 9+12. Te so v osnovnem kodu enake karakterjem +0, A, B, C, D, E, F, G, H, I; lahko pa pomenijo n.pr., da so vrednosti, ki so tako označene, nezaneslјive ali pa kakšno posebnost kodiranja.
3. V tej zadnji grupi so lahko karakterji večji ali enaki 41_8, a manjši ali enaki 52_8, to je enaki kombinacijam 0+11, 1+11, 2+11, 3+11, 4+11, 5+11, 6+11, 7+11, 8+11, 9+11. Te v osnovnem kodu pomenijo karakterje -0, J, K, L, M, N, O, P, Q, R, v našem kodu pa so to vrednosti, ki so večje od možnega zapisa veličine v pripadajočih kolonah, n.pr.

$\overset{x}{25}$ = K5 = 12,5 $\qquad$ $\overset{x}{00}$ = -00 = 100

Ta porazdelitev preseje podatke v tri različne grupe in izloči kot napačne vse vrednosti, ki niso tako označene. Tako izloči n.pr. kombinacijo $\overset{3}{8}$, katere pomen je nesmiseln, saj tretja in osma vrstica ne pomenita v osnovnem kodnem ključu numerično vrednost, niti ne karakter, ki bi padel v eno od zgornjih treh grup.

Stroj čita kartico postopoma, kolono za kolono, in posamezne podatke

razporedi v eno izmed treh dopustnih grup, če je podatek pravilen, sicer pa si v polju napak zapomni mesto, kjer je našel nepravilen karakter. Nato nadaljuje s čitanjem naslednje kolone po istem postopku vse do poslednje kolone na kartici. Čitanje kartice je tako končano. Sledi logična kontrola prečitanih podatkov, vendar le, če v polju napak ni nobene bistvene napake. Če pa je, se vsebina kartice izpiše skupno z oznako kolon, kjer so ugotovljene napake. Poleg tega pa se na luknjaču kartic napačna kartica duplira. Tako je tako kartico kasneje laže popraviti, saj je ni potrebno iskati v osnovnem paketu. Logična kontrola preveri vse kolone drugo za drugo, če so pravilno izpolnjene in v skladu z uporabljenimi kodami luknjanja kartic /4/. Na primer: $\overset{xx}{25}$ za temperaturo suhega termometra je napačno, saj drugi x ne pomeni ničesar. Pri postopku logične kontrole je potrebno še upoštevati, da lahko nastopajo spremembe koda.

Vrednosti prebrane s kartice se tako zberejo v polju podatkov in so pripravljene za posredovanje v glavni program. V polju podatkov je morebitni manjkajoči podatek dopolnjen s simbolično vrednostjo, n.pr. 99 ali 88, odvisno od opazovalnega obdobja in navodil.

Prikazano celotno metodo prenosa velikega števila podatkov s počasnih vhodnih medijev (kartic) na hitre (magnetne trakove) smo uporabili pri meteoroloških podatkih. Podatke - cca 10 milijonov - za 38 postaj Slovenije, ki so na 400.000 karticah /6/, smo prenesli v paketih dimenzij 31 krat 23 na magnetni trak. Prenos je tekel praktično brez zastojev. Edine težave so povzročale fizično slabe kartice (zmečkane ali zaradi slabega luknjača slabo izluknjane kartice). Celoten prenos je trajal 9 ur in je omogočil, da bo vse nadaljnje delo s temi podatki skoraj 100 hitrejše in preprostejše.

LITERATURA

/1/ Hočevar A. in Z. Petkovšek: Koncept kompleksne meteorološke obdelave z elektronskim računalnikom in nekaj rezultatov za meglo na letališču Ljubljana - Brnik. Razprave - Papers XI. DMS. Ljubljana 1969

/2/ Hočevar A.: Kontrola meteoroloških podatkov pri obdelavah z računalniki. Zbornik radova SHMZ. Beograd (v tisku)

/3/ Petkovšek Z.: Sodobno kompletiranje meteoroloških podatkov. Zbornik radova SHMZ. Beograd (v tisku)

/4/ Navodilo za vpisovanje klimatoloških podatkov v mehanografske obrazce. HMZ SRS. Ljubljana 1964 in 1966.

/5/ Computer Systems Fortran, Reference Manual, Central Data Corporation. November 1968. Pub. no. 60057600 C.

/6/ Arhiv klimatoloških podatkov SHMZ. Beograd.