

Ing. Svatoslav Andraška

ÚVT Ostrava

PARAMETRICKÝ PROGRAM TŘÍDĚNÍ A KUMULACE

Některé úvahy, související s problematikou třídění.

Třídění souborů dat je tak důležitou činností při hromadném zpracování dat, že samozřejmě součástí standardního programového vybavení každého počítače je generátor třídících programů, jehož vlastnosti redukuji úlohu programátora na zadání potřebných parametrů. V programovém vybavení počítače Tesla je to, jak známo, program TRI 140.

Zdálo by se tedy, že otázka třídění souborů dat je plně vyřešena a že není třeba se nadále tímto problémem zabývat. Ovšem v rámci hromadného zpracování dat je třídění tak často se opakující operací, která vzhledem k poměrně velké spotřebě strojového času /zvláště pak při třídění na magnetických páskách/ má podstatný vliv na ekonomiku a efektivnost vlastního zpracování informací, že je nutno hledat cesty, které tuto pomocnou operaci racionalizují.

V následujících řádcích chci poukázat na některé důležité poznatky /bez nároku na objevnost/, související s problematikou třídění, na jejichž základě byly formulovány požadavky na vlastnosti parametrického programu třídění a kumulace, který byl realizován v ÚVT Ostrava na počítači Tesla.

Jedním ze základních poznatků je, že při běžném a nejvíce rozšířeném způsobu používání programu TRI 140 má vý-

stupní seříděný soubor stejný počet vět, jako vstupní ne-seříděný soubor, tedy má stejnou úroveň rozlišení informací - což není vždy přirozeným pořadivkem.

Uplatníme-li v souborech informací rozlišovací úrovně, zjistíme, že jsou zejména výrazem účelu, pro který je soubor používán. V každém informačním subsystému jsou vytvářeny a periodicky doplňovány tzv. datové základny, které uchovávají informace na rozlišovací úrovni s nejvyšší podrobností rozlišení /a tudíž bývají velmi rozsáhlé/ a slouží jako základní soubory pro následný proces zpracování, přičemž proces zpracování a interpretace údajů se zprevidla děje na rozlišovacích úrovních s nižší podrobností rozlišení - podle účelu, pro který je tento proces určen.

Při analýze důvodů, vedoucích ke konkrétní specifikaci třídících hledisek /klíčů/, dojdeme k závěru, že právě třídícími klíči je dána rozlišovací úroveň následného zpracování a interpretace informací.

Věty souboru obvykle obsahují řadu položek, které lze rozdělit v každé fázi zpracování do tří skupin:

- položky, které větu jako logickou jednotku informací určují a nemají kvantitativní charakter /např. číslo závodu, číslo vozidla apod./. Podle těchto položek /klíčů/ se věty souboru např. třídí a volbou určitých položek této kategorie určujeme potřebnou rozlišovací úroveň,
- položky kvantitativního charakteru, jejichž obsah vyjadřuje výsledek jisté transakce /např. výnosy v Kčs apod./,
- položky, které jsou v průběhu zpracování v dané rozlišovací úrovni pasivní - někdy se jim říká výplňové.

Účelovou volbou třídících klíčů /a tedy i rozlišovací úrovně/ je výchozí soubor mnohdy rozčleněn na řadu skupin vět, které mají stejnou hodnotu klíčových položek. Obecně platí, že počet vět ve skupině roste se snižováním rozlišovací úrovně. V těchto případech je výhodné /před následným zpracováním/ provést kumulaci vět ve skupinách tak, aby ve výsledném souboru existovala ke každé hodnotě určených klíčů pouze jediná věta, v níž je hodnota kvantitativních polo-

žek součtem hodnot identických položek ze všech vět dané skupiny a výslednou hodnotou výplňových položek je hodnota z kterékoliv věty dané skupiny, anebo tyto mohou být vyplněny /pokud je to účelné/ domluvenými znaky. Tímto způsobem lze v mnoha případech dosáhnout podstatného fyzického zkrácení souborů pro následné zpracování při zachování účelové jednoty informací.

Při uvážení jediné možné organizace souborů na magnetických páskách, a to sakvenční organizace, je zřejmé, že předpokladem takto pojaté kumulace vět je příslušně seříděný soubor - což motivuje požadavek na úpravu algoritmu a rozšíření funkce třídícího programu tak, aby s procesem třídění byl spojen proces kumulace vět stejných hodnot klíčů.

Tímto algoritmem se dosáhne:

- v řadící fázi podstatného fyzického zkrácení monotónií a zmanženi jejich počtu /tento důsledek zvláště vynikne při částečně předtříděném vstupním souboru, což se v praxi stává dosti často/,
- ve slučovací fázi se tímto dosáhne podstatného snížení spotřeby strojového času a další kumulací vět mnohdy radikálního zmanženi objemu výstupního souboru.

Velmi častým požadavkem je výběr určitých informací z datové základny pro následné zpracování. I zde se ukazuje, že je výhodné dále rozšířit funkci třídícího programu tak, aby již během čtecí fáze mohl být prováděn příslušný výběr. Praktický přínos zařazení této funkce do algoritmu třídícího programu je nesporný.

Použití třídícího programu s uvedenými vlastnostmi je velice efektivní zejména v případě přípravy souborů pro následné tisky různých sestav evidenčních přehledů, statistických rozborů apod. Tyto sestavy jsou vesměs sumarizačního charakteru a různými počtem stupňů součtů, tvořených a řazených z různých hledisek. Při klasickém způsobu zpracování se po seřídění souboru podle potřebných klíčů /stupňů součtů/ provádí sumarizace hodnot za všechny stupně součtů přímo tiskovým programem. Použitím třídícího programu, který

dokáže kumulovat, se nejen podstatně zkrátí proces třídění tím, že vytvoření kumulovaného souboru se provádí sumarizací ze základní /nejnižší/ stupně součtí, ale rovněž se zkrátí celková doba tisku požadovaných sestav v důsledku toho, že tiskové programy zpracovávají soubory podstatně fyzicky zkrácené /tento důsledek obzvláště vynikne při používání obecného tiskového programu/.

V souvislosti s tiskem rozborových a evidenčních sestav vyslyne další požadavek na rozšíření funkce třídícího programu. V praxi se totiž velice často stává, že je nutno vytisknout sestavy, které jsou rozdílné jen rozlišovací úrovní, se které jsou zkoumány atributy jistých informačních elementů. Např. mají se vytisknout dvě sestavy, z nichž první je rozbor nákladů a výnosů v rámci hospodářských středisk a druhá sestava je rovněž rozbor nákladů a výnosů, ale již na úrovni závodu. Při klasickém způsobu zpracování se vždy před tiskem příslušné sestavy musí provést potřebné setřídění výchozího souboru. Tuto úlohu lze řešit úsporněji - pokud totiž třídící program umožní vytvoření duplikátu každé věty vstupního souboru, ve kterém bude "maskován" určitý klíč tak, že se automaticky sníží rozlišovací úroveň informací v těchto duplikátech, pak po setřídění a kumulaci bude výstupní soubor obsahovat jednak věty v požadovaném třídění /a rozlišovací úrovni/ pro tisk první sestavy a jednak věty v požadovaném třídění /a rozlišovací úrovni/ pro tisk druhé sestavy, a obě sestavy lze vytisknout jedním průchozím tiskovým programem.

Praktické důsledky, které lze očekávat při používání třídícího programu s popsányi možnostmi, jsou tedy zejména:

- podstatné zvýšení rychlosti třídění dat na magnetických páskách, a tím i snížení spotřeby strojového času při třídění,
- v souvislosti s kondenzací dat snížení spotřeby strojového času následných úloh /zvýšení ekonomiky návazného zpracování/,
- snížení spotřeby programátorské práce,

- standardizace šalších typů díloh ve zpracování dat.

Všechny uvedené funkce jsou realizovatelné parametrickým programem třídění a kumulace GTKP, který má i další výhodné vlastnosti - v souladu s potřebami, vyplynuvšími ze zkušeností s tvorbou a údržbou programových informačních subsystémů v ÚVT Ostrava.

Charakteristika programu GTKP.

GTKP je takový programový prostředek, který umožňuje rychle a s jednoduchou analýzou, bez nutnosti programování a ladění, na základě jednoduchého zadání a s minimálními zkušenostmi produkovat programy, které v sobě zahrnují /podle zvolených požadavků/ funkce výběru, třídění a kumulace vět informačních souborů na počítači Teal.

Vzhledem k tomu, že program musí splňovat požadavek na univerzálnost, tzn. že musí být schopen zpracovávat sekvenci souborů s různou strukturou a uložením dat ve větách buď pevného, nebo proměnného formátu, byla použita metoda parametrického řízení programu.

Parametry, jimiž jsou specifikovány funkce programu, jsou předávány programu buď ve formě, odpovídající operačnímu systému, a to prostřednictvím příkazových štítků, anebo pomocí speciálně vytvořené soustavy parametrických štítků, jejichž struktura a v určitém rozsahu i obsahová stránka odpovídá parametrickým štítkům, jimiž jsou definovány požadavky na funkce univerzálního programu tisku GPP nebo univerzálního programu pro práci se soubory GFLP, které jsou v ÚVT Ostrava již několik let široce využívány a proto bylo žádoucí dodržet jednotu formy parametrů.

Na základě zadaných parametrů se vytvoří v operační paměti modifikovaný program, schopný okamžitě plnit požadované funkce.

Parametry lze rozdělit do čtyř skupin:

1. Parametr, určující hlavní funkci programu
2. Parametry, charakterizující vstupní a výstupní soubor
3. Parametry, přiřazující souborům zařízení

4. Parametry, upřesňující v rámci hlavní funkce další požadavky na zpracování

ač 1/ Je možno zvolit jednu ze tří hlavních funkcí:

- prosté třídění vět /počet vět výstupního souboru je stejný, jako počet vět vstupujících do procesu třídění/,
- třídění se současnou kumulací vět na zvolenou rozlišovací úroveň,
- prostá kumulace vět na zvolenou rozlišovací úroveň /předpoklad vhodného setřídění vstupního souboru/.

ač 2,3/ Mají tentýž význam a způsob zadání jako u programu TRI 140. Určení minimální délky věty parametrem MINREC je povinné pouze u určitého formátu věty a navíc lze rozhodnout, zda magnetická páska výstupního souboru bude během třídění využita jako pracovní stopa, či nikoliv.

ač 4/ Tyto parametry jsou děrovány do děrných štítků, které tvoří vstupní soubor programu. Každý parametrický štítek této kategorie obsahuje ve sloupci 1-2 číselný kód, vyjadřující požadovanou funkci a od sedmého sloupce jsou uvedeny jednak údaje o poloze ve větě, se kterou bude operováno, jednak další potřebné a doplňující údaje. Vzhledem k tomu, že každá položka ve větě je určena:

- adresou vzhledem k počátku věty,
- délkou zóny, kterou ve větě zaujímá,
- formou /tvarem/ uložení v zóně,

je základním předpokladem pro věcně správný zázpis parametrů znalost struktury věty zpracovávaného souboru, přičemž se předpokládá, že odpovídající si položky jsou ve všech větách souboru umístěny na totéž místě, v totéž tvaru i rozsahu.

Prostřednictvím parametrických štítků této kategorie je umožněno zadání:

- klíče třídění, resp. kumulace
- položek, jejichž hodnota má být v rámci kumulace soustavně
- podmínek, na jejichž základě je prováděn výběr vět ze vstupního souboru, vstupujících do procesu třídění, resp. kumulace

- podmíněného vložení zvolených řetězců znaků nebo hexadecimálních čísel do předtčených vět vstupního souboru
- tvorby duplikátů vět, které jsou se zvolené modifikací jejich obsahu zařazeny do tříděného souboru vět
- vložení zvolených řetězců znaků nebo hexadecimálních čísel do všech vět výstupního souboru
- zákazu zápisu těch vět, ve kterých je výsledná hodnota všech sumarizovaných položek současně nulová
- kontroly počtu skutečně sejmutých parametrických štítků této kategorie proti kontrolnímu součtu.

Všechny zadané parametrické údaje jsou opsány na tiskárně a jsou důkladně kontrolovány jak z hlediska formátu, tak i obsahu a případné chyby jsou indikovány upozorněními na referenčním osacím stroji a detailními popisy chyb na tiskárně. Systém je dále doplněn provozními zprávami na referenčním osacím stroji a na tiskárně tak, aby obsluha i uživatel byli řádně informováni o průběhu a výsledku zpracování.

Při konstrukci programu byl kladen důraz na optimální využití disponibilní operační paměti.

Pro účinnost uvádím, že v programu GTRP byl s výhodou aplikován třídící podprogram, který byl vyvinut ve výpočetním středisku Tesly Rožnov - za prvé proto, že při jeho vývoji bylo namatováno na možnost zpracování vět se shodnými klíči /mj. tím, že provozní rutina musí naprogramovat uživatel/ a za druhé a to především, poněvadž použitý algoritmus vnitřního třídění /modifikace tzv. stromáčku/ umožňuje vytvářet monotónie, jejichž délka není shora omezena velikostí pracovní oblasti paměti, což zajišťuje celkově lepší třídící časy než je tomu u programu TRI 140.

Shodně s přínosu programu GTRP pro racionalizaci a zvýšení ekonomiky zpracování některých úloh na počítači Tesla.

Činnost důsledně tvorby účelově kumulovaných souborů /a v takovémto případě lze už hovořit o metodě/ byla poté, kdy pro racionální uplatnění této tvorby byl k dispozici

podpůrný programový prostředek, ověřena v subsystému Komplexní účetnictví. Při programové realizaci tohoto projektu, který v současné podobě zahrnuje výstup více než 80 uživatelských sestav, bylo v převážné míře použito parametrického tiskového programu GPP, což mimo jiné umožnilo poměrně rychlé zavedení tohoto subsystému do rutinního zpracování. Velice zjednodušeně řečeno, celé zpracování spočívá /po vytvoření příslušných datových základů/ v neustálém třídění a tisku požadovaných sestav.

Výsledky tohoto ověření plně potvrdily naše předpoklady ve značném snížení spotřeby strojového času a tím, že výhodnější předzpracování tiskových souborů eliminovalo určitý nedostatek programu GPP /spočívající v nevhodném rozvrhování součtových řádků na stránce/, bylo dosaženo i úspory papíru a v neposlední řadě i zvýšení spokojenosti odběratelů s tvarem sestav.

Programem GTPK byla vhodně doplněna soustava parametrických programů, které v ÚVT Ostrava používáme a v současné době jsme schopni "na počkání" splnit požadavky na mimořádné zpracování i takových sestav, jejichž tisk bychom museli dříve naprogramovat.

Naop. vznikne požadavek na zpracování sestavy, která má být podnikovým přehledem vzniklých nákladů na jednotlivých objektech /kalkulačních jednicích/ v typovém členění podle kalkulačního vzorce, přičemž formát sestavy určuje, že v jednom řádku bude uveden jeden objekt s příslušnými složkami nákladů. V datové základně jsou informace uchovány tak, že každá věta obsahuje vždy hodnotu jen jedné možné složky nákladů, a to v podrobnějším analytickém členění. Vzhledem k tomu, že program GPP dokáže tisknout pouze vertikální součty, nemohl by být pro tisk požadované sestavy přímo použit. Kyní tuto úlohu zvládneme za použití pouze parametrických programů, které máme k dispozici, přičemž postup spočívá v tom, že se nejprve vytváří soubor, v jehož větách zaujmají hodnoty příslušných složek nákladů pevně definované pozice /navzájem rozdílné/, poté se provede požadované třídění se sou-

řádnou kumulací a vzniklý soubor umožní programu GPP položkově tisknout sestavu s horizontálně umístěnými součty.

Jako další příklad racionalizace lze uvést tuto úlohu - v souboru, který je setříděn podle jistých klíží a je kumulovaný, je zapotřebí provést opravu chybných hodnot některých klíčových položek, vyskytujících se v několika větách. Obecně je nutno předpokládat, že po provedené opravě může být porušena dané setřídění a může vzniknout více vět se shodnými klíži. V takovém případě bylo nutno tuto úlohu řešit ve třech krocích s vyvoláváním jednotlivých programů /oprava, třídění a kumulace/. Programem GTRP lze tuto úlohu vyřešit najednou, což je samozřejmě ve všech směrech výhodnější.

Daly by se uvést ještě další příklady úloh, jejichž řešení je právě existencí vhodného podobného prostředku už jistým způsobem standardizováno, ale vymezený rozsah tohoto příspěvku je v tomto případě limitujícím faktorem.

Záměrem tohoto příspěvku bylo ukázat, že i tak samozřejměnou činnost ve zpracování dat, jako je třídění, lze při respektování jistých souvislostí racionalizovat, a tím přispět ke zvýšení ekonomiky zpracování a úspoře programátorské kapacity.

Literatura

- V.Šulc : Rozšíření funkce generátoru třídících programů.
Mechanizace automatizace administrativy, 1975
/KV/, 12, 494 - 496