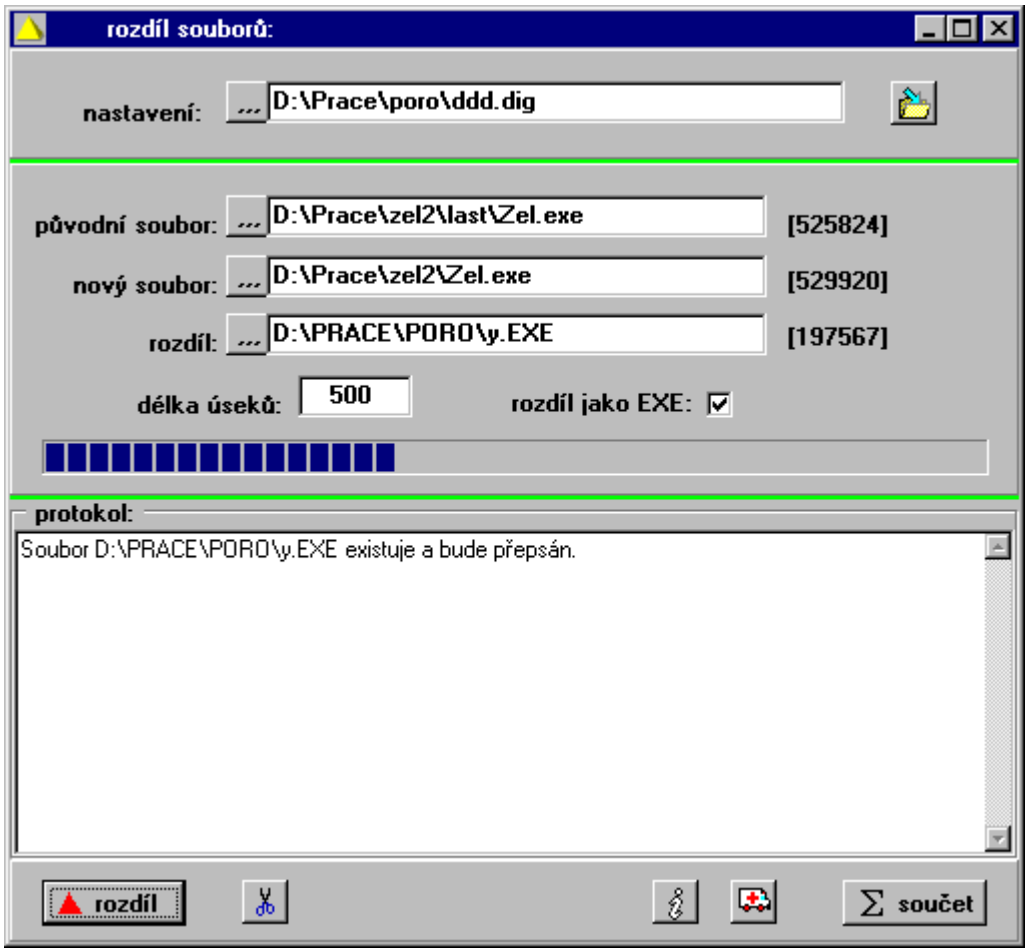
Popis programového systému DELTA.

Smysl programu vychází z následující situace. Potřebuji dopravit na místo určení novou verzi rozsáhlého (datového nebo exe) souboru. V souboru došlo oproti předchozí verzi k řadě změn. Pokud příjemce má k dispozici předchozí verzi souboru, je většinou výhodnější poslat informaci o tom, jak z původního souboru „poskládat“ novou verzi, než dopravovat celý soubor nový. Informace o sestavení nového souboru ze starého může být podle konkrétní situace i velmi podstatně kratší než nový soubor.

Program DELTA dokáže porovnat dva soubory - „původní“ s „novým“ – a vytvořit soubor informací o rozdílech a shodných úsecích (diferenční soubor). Výsledný soubor může být vytvořen ve dvou různých variantách a to buď jako datový, nebo jako EXE soubor.

Znovuvytvoření nového souboru je možné provést několika způsoby:
- aplikací programu DELTA, který umí provádět i zpětný proces znovuvytvoření nového souboru,
- aplikací programu SUMA
- spuštěním diferenčního souboru vytvořeného ve tvaru EXE souboru.

V prvním i v druhém případě lze pro znovuvytvoření použít diferenční soubor jak ve tvaru datového souboru tak ve tvaru EXE souboru.

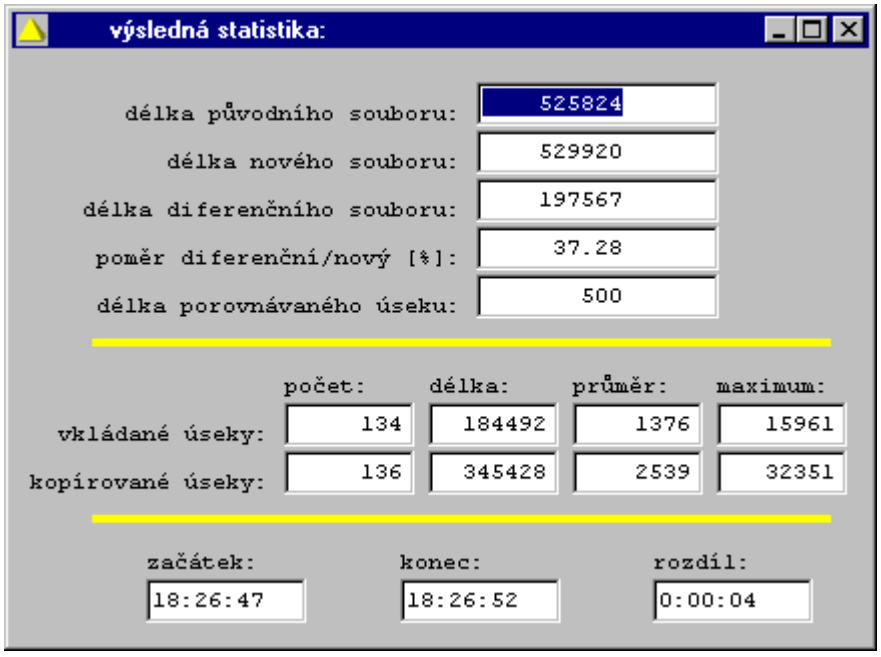


Obr. 1 – Vytvoření diferenčního souboru.

Program DELTA

Vytvoření diferenčního souboru

- Základními editačními poli jsou „původní soubor“, „nový soubor“ a „rozdíl“. Jejich smysl je zřejmý z názvu. Hodnoty je možno zadat ručně nebo poklepáním vyvolat klasický dialog pro výběr souboru.
- Pole „délka úseků“ určuje minimální délku vyhledávaných shodných úseků. Vzhledem k tomu, že informace o délce shodného úseku uložená do diferenčního souboru má 8 znaků, nemá smysl aby tato hodnota byla menší než 10. Zadaná délka má kromě toho i vliv na rychlost zpracování – čím kratší délka úseku, tím delší doba zpracování. V praxi má smysl zadávat tuto hodnotu od 100 u menších souborů (100-500K) po 300 až 500 u větších souborů.
- Zaškrtávací pole „rozdíl jako exe“ určuje formu vytvořeného rozdílového souboru – jako datový nebo jako EXE soubor.
- V poli „nastavení“ je možno stejně jako u výše uvedených polí buď ručně nebo poklepáním zadat jméno souboru, z něhož se načtou (resp. se do něho uloží po stisknutí následujícího tlačítka) všechny výše uvedené informace. Pokud tedy budete vytvářet diferenční soubory pro nějakou kombinaci souborů častěji, vyplatí se informaci o nastavení uložit a pak nastavení vyvolat zadáním tohoto souboru.
- Proces vytvoření diferenčního souboru se spustí tlačítkem „rozdíl“. Po jeho stisknutí se zkontroluje zadání a poté se spustí vytváření diferenčního souboru. Současně se na spodní liště objeví tlačítko s nůžkami. Jeho stisknutím můžete proces přerušit. Po skončení akce se pak objeví jiné tlačítko s kalkulačkou a po jeho stisknutí se zobrazí panel s detailními informacemi o vytvořeném diferenčním souboru (viz obr. 2).



Obr. 2 – Statistika po vytvoření diferenčního souboru.

- Do diferenčního souboru jsou uloženy informace o CRC (32 bit) původního a nového souboru. Tyto informace slouží při znovuvytváření souboru jednak pro kontrolu správnosti zadaného původního souboru a jednak pro kontrolu správnosti vytvořeného nového souboru.

- Po vytvoření diferenčního souboru se pro kontrolu ihned provede i zpětný proces znovuvytvoření nového souboru.

Zpětné vytvoření nového souboru

- Pro zpětné vytvoření nového souboru stačí zadat jméno „původní soubor“ a „rozdíl“. Hodnoty je možno zadat ručně nebo poklepáním vyvolat klasický dialog pro výběr souboru. Při tomto zadání se po vytvoření nového souboru přejmenuje starý soubor na soubor s příponou .OLD a nový soubor se bude jmenovat jako starý soubor. Pokud bude zadáno i jméno „nový soubor“, jsou ponechána jména dle zadání.
- Vlastní proces vytvoření nového souboru proběhne po stisknutí tlačítka s textem „součet“. Přitom se zkontrolují CRC hodnoty původního i nového souboru.

Program SUMA

Program SUMA plní stejnou funkci jako výše uvedené tlačítko v programu DELTA. Rozdíl je jen v tom, že SUMA je DOSovský program a parametry jsou zadávány z příkazového řádku. Volání se provádí příkazem:

SUMA starý_soubor diferenční_soubor [nový_soubor],

kde význam parametrů je opět zřejmý. Pokud není zadáno jméno nového souboru, bude po procesu vytvoření starému souboru přidělena přípona .OLD a nový soubor se bude jmenovat stejně jako starý soubor.

Například:

SUMA DATA.TXT DATA.DIF NOVADATA.TXT

Pomocí souboru DATA.DIF se z původního souboru DATA.TXT vytvoří nový soubor NOVADATA.TXT.

SUMA DATA.TXT DATA.DIF

Pomocí souboru DATA.DIF se z původního souboru DATA.TXT vytvoří nový soubor, který se bude rovněž jmenovat DATA.TXT, a starý soubor se přejmenuje na DATA.OLD.

Diferenční soubor jako EXE program

Tento program plní stejnou funkci jako program SUMA. Jediný rozdíl je v tom, že není nutné zadávat druhý parametr, kterým je název diferenčního souboru, protože tato data jsou obsažena přímo ve vytvořeném programu. Spuštění se tak provádí:

Jméno_programu starý_soubor [nový_soubor]

Současná omezení:

Maximální počet nalezených shodných intervalů může být nejvýše 1000.

Obr. 3 – Logo.

Poznámky k algoritmu.

Pokud by program vyhledával shodné úseky „byte po byte“ byla by doba zpracování již v případě nepříliš velkých programů neúměrně dlouhá. Program DELTA proto využívá pro urychlení práce různé heuristiky. Tím ale může nastat situace, že nebudou nalezeny zcela všechny vhodné shodné úseky. Základní princip práce je následující. Program vezme úsek zadané velikosti z „nového“ souboru a hledá, zda se vyskytuje v souboru „původním“.

- Pokud takový úsek nalezne, pokusí se tento úsek rozšířit o maximum dalších shodných znaků na obě strany – na začátku i na konci úseku. Výslednou informaci si poznačí a nový úsek zadané velikosti začíná v „novém“ souboru na následující pozici za takto nalezeným úsekem.
- Pokud není takový úsek nalezen, vezme se další úsek zadané velikosti následující za tímto úsekem.

Po skončení vyhledávání se vytvoří výstupní „diferenční“ soubor tak, že se (za informace o CRC) zapíší informace o shodných úsecích, doplněné o sekvence znaků, které v původním souboru nalezeny nebyly.

Matematicky vzato, je-li M zadaná délka úseku a N je délka shodného úseku v „novém“ souboru, pak pravděpodobnost (p), že právě tento úsek se bude vyhledávat v „původním“ souboru je

- p=0 pro N<M,
- p= (M-N+1)/M pro 2*M>N>M,
- p=1 pro N>=2*M.

Čili není otázkou zda se shodný úsek v původním souboru nalezne (pokud tam existuje pak se určitě nalezne), ale otázkou je, zda se k vyhledávání vybere. Výše uvedené vzorečky tedy říkají, že je-li shodný úsek menší než zadaný, pak se určitě nevybere

k vyhledání, je-li dvojnásobně velký nebo větší, pak se určitě vybere (a tedy i nalezne) a v případě, že je mezi M a 2*M pak je vybrán k vyhledávání s výše uvedenou pravděpodobností.

Jiří Ventluka,
jive@vol.cz

Praha, 10. 3. 1999