# Indexování obsahu souborů

Miroslav Hrončok

5. července 2011

## 1 Zadání

Mým úkolem vylo vytvořit spolupracující dvojici programů. Jeden bude indexovat obsah souborů a druhý bude využívat výsledky práce toho prvního. Vyhledání slova ve větším množství dat je časově náročné, a proto je tento způsob efektivní. Celá koncepce funguje podobně jako linuxové programy updatedb a locate, které ale indexují pouze názvy souborů. Můj program indexuje obsažená slova.

## 2 Vlastnosti

Stanovil jsem si některé cíle, které by měl program splňovat z pohledu uživatele:

### 2.1 POSIX

Program by měl bez problému fungovat na operačním systému posixového typu. Testován byl však pouze v GNU/Linuxu.

### 2.2 Variabilita

Program umožňuje jeho uživateli určitou variabilitu. Je možno se rozhodnout, který adresář chce indexovat, zda indexovat i skryté soubory a složky, zda záleží na velikosti písmen, či zda čísla figurují jako součást slov, nebo jako jejich separátory. Zároveň si uživatele může vybrat ze dvou párů programů. Ty, které začínají písmenem e, prohledávají soubory dle regulárních výrazů (jeden výraz ale může pokrýt pouze jedno slovo) a ty, které začínají písmenem w, prohledávají soubory pouze podle celých slov. Oba páry programů jsou optimalizované pro konkrétní užití.

### 2.3 Indexace změn

Po první indexaci zvoleného adresáře se indexují pouze změny. Další indexace jsou tedy znatelně rychlejší, než ta prvotní.

## 3 Programování

Kromě cílů, které pocítí uživatel, jsem se snažil dodržet i některá pravidla jako polymorfismus, zapouzdření, modularita apod.

### 3.1 Jak to funguje

Princip indexace je v zásadě jednoduchý:

1. Načte se seznam minule indexovaných souborů a jejich časů úpravy při minulé indexaci
2. Načte se seznam reálných souborů a případné rozdíly se promítnou do již načtených dat
3. Pokud došlo ke změnám, načte se seznam slov obsahující informaci, ve kterých souborech se dané slova nacházejí
4. Smažou se data z neexistujících a starých souborů
5. Nově vytvořené a upravené soubory se zindexují
6. Slova a nový seznam souborů se uloží do skryté složky v indexovaném adresáři

Ve variantě w je každé slovo je reprezentováno jedním souborem, kde název souboru odpovídá slovu a obsah souboru je seznamem souborů, které dané slovo obsahují. Vyhledání slova pak předupravuje pouze otevření konkrétního souboru. Ve variantě e jsou pak jednotlivá slova uložena do jednoho souboru, protože se všechna indexovaná slova stejně musejí porovnat s regulárním výrazem.

Pro pochopení vztahů mezi třídami jsem dle požadavku vytvořil schma, které najdete v souboru *classes.eps*.

## 4 Omezení

Program samozřejmě není všemocný a má nějaká omezení. Indexuje pouze ASCII textové soubory, takže s háčky a čárkami si neporadí. Vyhledávání pomocí regulárních výrazů je navíc omezeno pouze na jedno slovo – nelze vyhledávat sousloví.

## 5 Test

Pro test jsem zvolil zdrojové kódy uživatelského prostředí Xfce 4.8 [http://wiki.xfce.org/dev/howto/git]. Jejich celková velikost je 151,4 MB a obsahují celkem 8276 položek (adresářů a souborů). První indexace trvala u obou variant necelých 8 minut. Samotné vyhledání z podstaty proběhlo prakticky okamžitě. V adresáři env jsou testovací soubory, které jsem používal během ladění (obsahují různé nástrahy), můžete je použít pro časově méně náročný test. Před použitím si přečtěte README nebo zavolejte program edbzer nebo wdbzer s argumentem –help, jinak bude indexovat vaši domovskou složku, což může trval v závislosti na její velikosti velmi dlouho.

Pokud budete testovat regulární výrazy, nezapomeňte argument programu uvést v jednoduchých uvozovkách, aby interpret řádky nespolkl speciální znaky.