Licence d'informatique
Module de C/Unix

Projet Unix

`stfs` — Search Though FileS

Philippe MARQUET

Octobre 1997

La commande `stfs` (*Search Though FileS*) proposée ici est inspirée de la commande `glimpse` développée par Udi MANDER, Sun WU, et Burra GOPAL, University of Arizona, Tuscon, AZ (`http://glimpse.cs.arizona.edu/`).

La commande `stfs` sera développée en Bourne Shell (`sh`).

Recherche d'informations

La recherche d'informations dans un système de fichiers Unix recouvre l'utilisation de plusieurs outils. Les commandes de recherche de type `grep` sont rapides si la recherche se limite à un domaine restreint (fichiers peu nombreux et peu volumineux). La recherche à l'aide d'outils basés sur des index nécessite la construction de fichiers d'index volumineux (typiquement de l'ordre de 50 à 300% des informations indexées).

La commande `stfs` est une combinaison de ces deux approches. L'utilisation de cette commande est ainsi possible même pour des recherches dans des systèmes de fichiers personnels pour lesquels l'utilisateur ne peut pas doubler l'utilisation de son espace disque par la création de volumineux fichiers d'index.

La méthode de recherche utilisée au sein de `stfs` est une *recherche à deux niveaux*. La mise en place de cette méthode est motivée par le fait que les recherches séquentielles telles que celle effectuées par `grep` sont efficaces pour des recherches dans des textes de taille raisonnable. Il n'est donc pas nécessaire d'indexer exactement l'ensemble des mots. L'approche de la recherche à deux niveaux est de ne pas mémoriser dans un fichier d'index l'emplacement exact des mots mais un pointeur vers une zone dans laquelle la réponse précise sera trouvée. Ensuite, une recherche séquentielle de type `grep` sera utilisée dans cette zone pour présenter la solution exacte.

Construction de l'index

L'espace d'information est un ensemble de fichiers textes. Un texte est consisté d'une séquence de *mots* séparés par les délimiteurs habituels (espace, fin de ligne, point, virgule...).

La première phase de l'indexation est de diviser l'espace d'information en *blocs*. Cette division est faite de manière à ce que les blocs soient de taille équivalente. De manière générale, le nombre de blocs doit être inférieur à une valeur donnée, en effet les blocs seront adressés par une clé qui sera conservée dans le fichier d'index. Afin de faciliter les choses dans le cadre de notre projet développé en shell, les clés seront représentées par un caractère. Les caractères imprimables étant limités à 94 caractères, nous nous limiterons à 94 blocs.

L'ensemble de l'espace d'information est analysé mot par mot pour construire un index. À l'inverse des index habituels dans lesquels chaque occurrence d'un mot est indexée par un pointeur sur sa position exacte, dans notre approche tous les mots sont indexés, mais pas chacune des occurrences. Une entrée dans l'index contient un mot et la liste des blocs dans lesquels le mot apparaît. Même si un mot apparaît plusieurs fois dans un bloc, seul l'index du bloc apparaît dans l'index, et une seule fois.

Id : stfs.tex,v 1.5 2000/12/03 10 :47 :49 marquet Exp

Recherche d'un mot

La recherche est faite en deux phases. Premièrement, on réalise une recherche dans l'index pour identifier l'ensemble des blocs qui peuvent satisfaire la requête. Ensuite une recherche dans ces blocs est faite de manière indépendante.

La recherche dans l'index est faite de manière séquentielle avec `grep` et non en utilisant des algorithmes de hashage. L'avantage d'une telle approche est la possibilité d'utiliser les expressions régulières de `grep` et donc de ne pas seulement rechercher des mots au sens strict du terme.

Syntaxe de la commande `stfsindex`

la commande `stfsindex` crée un index pour les fichiers textes d'une arborescence dont le nom de la racine est passé en paramètre (par défaut, `stfsindex` indexe le répertoire `$HOME` de l'utilisateur) :

`stfsindex` [`-z`] [`-l`] [*dirname*]

L'option `-z` indique à `stfsindex` de considérer les fichiers textes et les fichiers textes compressés.

L'option `-l` signale de ranger dans un fichier de log l'évolution de l'exécution de la commande (nom des fichiers traités, liste des fichiers n'ayant pu être indexés suite à une erreur...) afin de pouvoir localiser d'éventuelles erreurs.

Implémentation de la commande `stfsindex`

Préalablement à l'indexation d'un fichier, la commande `stfsindex` doit déterminer si le fichier est un fichier texte. Elle utilise pour cela la commande `file (1)`.

Si l'option `-z` est présente, les fichiers compressés ou codés sont décompressés/décodés et considérés comme des fichiers textes si le résultat de la décompression est un fichier texte. Les utilitaires de décompression utilisés sont `uncompress (1)` et `gunzip (1)`. L'utilitaire de décodage considéré est `uudecode (1)`.

La partition des fichiers à indexer en blocs peut être réalisée de la manière suivante. La somme des tailles des fichiers à indexer est calculée ; une taille moyenne des blocs en découle. Les fichiers sont combinés jusqu'à atteindre cette taille et un nouveau bloc est commencé.

Le résultat de l'indexation doit fournir :
- la liste des fichiers correspondant à un bloc donné ;
- l'index lui-même.

Ces informations seront rangées dans des fichiers `.stfs.*` situés dans le répertoire de login de l'utilisateur. Il est possible d'organiser ces informations de multiples manières. Nous précisons ici une de ces possibilités.

Les informations sont réparties dans les fichiers :

`$HOME/.stfs.index` est structuré en lignes. Une ligne débute par un mot suivi d'un séparateur (un espace) et de la liste des codes (caractères) des blocs dans lesquels le mot apparaît.

`$HOME/.stfs.filenames` contient la liste des fichiers indexés à raison d'un nom (absolu) de fichier par ligne. Un caractère supplémentaire code le fait que le fichier est compressé/codé.

`$HOME/.stfs.blocks` définit les ensembles de fichiers appartenant à un bloc donné à raison d'un bloc par ligne. Ce fichier contient (au plus) autant de lignes que de valeur différentes pouvant être prise par l'index d'un bloc (c'est-à-dire pour ce qui nous concerne 94 lignes). La ligne i correspond au i^{e}bloc. Elle précise l'intervalle des fichiers appartenant au bloc. Les numéros des fichiers correspondent aux numéros de ligne dans le fichier `$HOME/.stfs.filenames`.

`$HOME/.stfs.info` donne l'ensemble des informations associées à l'exécution de la commande `stfsindex` : date de création de l'index, nom de l'arborescence indexée, nombre de fichiers indexés, nombre de fichiers non-indexés de l'arborescence, taille moyenne des blocs, temps d'exécution de la commande `stfsindex`, prise en compte ou non des fichiers compressés/codés, etc.

`$HOME/.stfs.log` est le fichier de log créé par l'option `-l` de la commande `stfsindex`.

Syntaxe de la commande `stfs`

La commande `stfs` recherche les lignes des fichiers indexés par `stfsindex` contenant un *mot* qui corresponde à l'expression régulière *pattern* passée en paramètre. Le format d'affichage est équivalent à celui de `grep` : la ligne est précédée du nom fichier. La syntaxe de la commande est :

`stfs [options]` *pattern*

Les options de `stfs` sont un sous-ensemble des options de `grep` ; elles incluent au moins les options

`-i` Ignore la distinction entre les majuscules et les minuscules lors des comparaisons.

`-l` Affiche uniquement les noms des fichiers à raison d'une ligne par fichier. Si un *pattern* est trouvé plusieurs fois dans un fichier, le nom du fichier n'est affiché qu'une fois.

`-c` N'affiche que le nombre de lignes qui correspondent au *pattern*.

`-q` Quiet. Ne produit pas de résultat sur la sortie standard quelque soit le résultat de la recherche. La commande retourne une valeur nulle si et seulement si une ligne au moins correspond au *pattern*.

La commande `stfs` retourne

0 si une ou plusieurs occurrences du *pattern* ont été trouvées ;

1 si aucune occurrence n'a été trouvée ;

2 si les fichiers d'index n'ont pas été préalablement construits à l'aide de la commande `stfsindex;`

3 en cas d'erreur d'accès aux fichiers de données (même si des occurrences du *pattern* ont été trouvées) ;

Extensions

- Si il n'est pas envisageable d'indexer des fichiers qui ne contiennent pas du texte (par exemple les exécutables), certains fichiers textes n'ont pas à être indexés. Un bon exemple est un fichier contenant des séquences d'ADN. Il est alors possible de spécifier qu'un fichier ne doit pas être indexé en ajoutant son nom à un fichier `$HOME/.stfs.prohibit` ou à un fichier `.stfs.prohibit` dans le répertoire du fichier.
- Donner la possibilité à l'utilisateur de contrôler la partition des fichiers en blocs.
- Dans le cadre du projet de C associé : développement d'une commande de type `grep` spécifique. On peut par exemple introduire des ¡¡ et ¿¿ qui seraient gérés au niveau des index : les blocs à scruter seraient ceux dans lesquels apparaissent les deux patterns.

Outils

La réalisation des commandes `stfs` et `stfsindex` nécessite la conversion d'un index de bloc en un caractère imprimable. Les commandes suivantes, écrites en C, sont données ; on trouvera les exécutables dans le répertoire `~marquet/c-unix/stfs` :

`index2char` accepte en paramètre un index de bloc (valeur comprise entre 1 et 94) et affiche sur la sortie standard un caractère correspondant au codage de cet index.

`char2index` accepte en paramètre un caractère et retourne un index de bloc ; valeur entière comprises entre 1 et 94.

Ces deux commandes retournent une valeur non nulle en cas d'erreur.

Réalisation

La première chose à faire est de relire plusieurs fois en détail le présent sujet !

Vous devez rendre (sous une forme à définir avec votre enseignant de TD/TP) :

– un compte-rendu de projet décrivant les commandes développées, les algorithmes mis en place et la structure des fichiers d'index ;
– les sources shell des commandes développées ;
– un test de ce projet sous la forme de résultats d'exécution des commandes. En particulier on comparera les temps d'exécution de la commande `stfs` et d'une recherche directe par `grep`.