Ημερομηνία παράδοσης: 18/02/2011

Εργασία 1

Εισαγωγή:

Η αναζήτηση πληροφοριών στον Παγκόσμιο Ιστό (Web) είναι μια πολύ σημαντική λειτουργία που συμβάλλει σε μεγάλο βαθμό στην αύξηση της ευχρηστίας του. Οι περισσότεροι χρήστες του Web χρησιμοποιούν μηχανές αναζήτησης για να ανακαλύπτουν πληροφορίες (μέσω ιστοσελίδων) που χρειάζονται για τις ανάγκες τους. Η βασική τεχνική που χρησιμοποιήθηκε από τις μηχανές αναζήτησης για την ανακάλυψη των πληροφοριών αυτών είναι το Web crawling.

Οι Web crawlers είναι προγράμματα υπολογιστών που ψάχνουν το Web με μεθοδικό και αυτοματοποιημένο τρόπο, ανακαλύπτοντας πληροφορίες που βρίσκονται σε διάφορες ιστοσελίδες.

Εκφώνηση Εργασίας:

Σκοπός αυτής της εργασίας είναι να δημιουργήσετε ένα πολύ απλό crawler. Η διαφορά από ένα πραγματικό Web crawler είναι ότι αντί ιστοσελίδες, ο δικός σας crawler θα ανακαλύπτει αρχεία (files) και τα περιεχόμενα τους.

Η λειτουργία του crawler θα έχει ως εξής:

1. Ο crawler θα λειτουργά ως ένας απλός πελάτης (client) και θα ψάχνει για εξυπηρετητές (servers) που περιέχουν κάποια αρχεία.
2. Ο crawler αυτός θα πρέπει να λαμβάνει από κάποιο τοπικό (local) αρχείο τις διευθύνσεις των εξυπηρετητών (που περιέχουν τα αρχεία) και θα επικοινωνεί μαζί τους μέσω του Web.
3. Σε πρώτο στάδιο, η επικοινωνία θα περιλαμβάνει αναζήτηση στον εξυπηρετητή ενός αρχείου με όνομα index.txt. Στο αρχείο αυτό θα υπάρχουν καταγεγραμμένα τα αρχεία που υπάρχουν στον εξυπηρετητή.
4. Αφού ο crawler πάρει τα περιεχόμενα του αρχείου index.txt, θα πρέπει να επικοινωνεί διαδοχικά με τον εξυπηρετητή για να πάρει τα περιεχόμενα των αρχείων που περιέχονται στο αρχείο index.txt.
5. Αφού πάρει τα περιεχόμενα των αρχείων, θα πρέπει απλά να τα τυπώνει στην οθόνη, το ένα μετά το άλλο.

Σημειώσεις:

- Για τους σκοπούς της άσκησης θα πρέπει να γράψετε κώδικα για ένα απλό εξυπηρετητή και ένα απλό πελάτη (crawler), οι οποίοι θα μπορούν να επικοινωνούν μέσω του ιστού.
- Για σκοπούς απλοποίησης, σε πρώτο στάδιο μπορείτε να θεωρήσετε ότι το αρχείο με τις διευθύνσεις των εξυπηρετητών περιέχει μόνο ένα εξυπηρετητή. Αυτός μπορεί να είναι η ίδια μηχανή (local host) ή κάποιος απομακρυσμένος εξυπηρετητής (remote server) στον οποίο θα τρέξετε τον κώδικα του εξυπηρετητή σας (π.χ. στον λογαριασμό σας σε κάποια μηχανή του εργαστηρίου).
- Ο εξυπηρετητής θα πρέπει να δέχεται απλές εντολές σαν αιτήματα από τον πελάτη-crawler. Σε αυτή τη φάση της εργασίας πρέπει να δέχεται απλώς μια εντολή (GET index.txt) που να επιστρέφει μια λίστα με τα αντικείμενα-αρχεία που προσφέρει και μια εντολή (GET {file}) που επιστρέφει κάποιο αρχείο.
- Το πρωτόκολλο επικοινωνίας θα πρέπει είναι στο επίπεδο του TCP/IP μέσω socket programming και όχι στο επίπεδο του HTTP.

Γενικές Πληροφορίες:

Για τη συγγραφή του προγράμματος θα χρησιμοποιηθεί η γλώσσα Java ή η γλώσσα C. Σε περίπτωση που επιλέξετε την γλώσσα C, θα πρέπει να βεβαιωθείτε ότι η εφαρμογή σας μπορεί να τρέξει στα εργαστήρια B103 (για σκοπούς συμβατότητας με θέματα compilation). Η ημερομηνία παράδοσης ορίζεται στις 18 Φεβρουαρίου 2011.

Τα παραδοτέα τις εργασίας αυτής είναι:

- Ο πηγαίος κώδικας του προγράμματος (source code)
- Μια μικρή ανάλυση και τεκμηρίωση (documentation) της υλοποίησης σας και πιθανά σχόλια για δυσκολίες που αντιμετωπίσατε (1-2 σελίδες το μέγιστο)

Τα παραδοτέα θα πρέπει να αποσταλούν με email στον βοηθό του μαθήματος Αντρέα Καμηλάρη (email: kami@cs.ucy.ac.cy), επισυναπτόμενα (attached) σε μορφή zip file.