ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ
ΗΥ 463 – Συστήματα Ανάκτησης Πληροφοριών

Εργασία: Ταραντούλα

Εισαγωγή

Σκοπός της εργασίας αυτής είναι η σχεδίαση και η ανάπτυξη ενός **ερπυστή** σε Java που να καλύπτει τις λειτουργικές ανάγκες του Mitos. Οι ερπυστές χρησιμοποιούνται κυρίως από μηχανές αναζήτησης για το «κατέβασμα» ιστοσελίδων. Οι επιθυμητές σελίδες καθορίζονται δίδοντας ένα σύνολο URLs εκκίνησης και προσδιορίζοντας την επιθυμητή πολιτική διάσχισης και σταχυολόγησης. Συνοπτικά, ένας ερπυστής λαμβάνει ως είσοδο ένα αρχείο παραμετροποίησης (Configuration file) και παράγει ως έξοδο έναν τοπικό κατάλογο που περιέχει αντίγραφα των ιστοσελίδων που επισκέφτηκε ο ερπυστής και ένα αρχείο-ευρετήριο (index-file) με πληροφορίες για αυτά τα αντίγραφα. Γενικές πληροφορίες για τους ερπυστές θα διδαχθούν στο μάθημα και υπάρχουν στα βιβλία του μαθήματος. Αναλυτικότερες οδηγίες δίδονται στην επόμενη ενότητα.

Η υλοποίηση του ερπυστή πρέπει να είναι παραμετρική και εύκολα επεκτάσιμη. Για το λόγο αυτό συστήνεται η δημιουργία UML διαγραμμάτων που να περιγράφουν την σχεδίαση του ερπυστή. Συνάμα ο κώδικας σας πρέπει να είναι πλήρως τεκμηριωμένος με σχόλια και Javadocs. Τέλος για το έλεγχο της εγκυρότητας του κώδικα σας συστήνεται η χρήση του JUnit.

Για να βοηθηθείτε (στη σχεδίαση αλλά και στην υλοποίηση) μπορείτε να επισκεφθείτε τους παρακάτω συνδέσμους:

- Παράδειγμα ενός απλοϊκού ερπυστή:
  http://java.sun.com/developer/technicalArticles/ThirdParty/WebCrawler/
- Ερπυστές ανοιχτού / ελεύθερου κώδικα:
  http://java-source.net/open-source/crawlers
- Επεξήγηση των περιεχομένων των http-επικεφαλίδων:
  http://www.w3.org/Protocols/rfc2616/rfc2616-sec14.html

Γενικός Αλγόριθμος

Ακολουθεί μια γενική περιγραφή του αλγορίθμου ενός ερπυστή.

```
Read the configuration file
Q =  Seeds  // a set of URLs  (the starting points)
While ( not(Empty(Q)) and  Not(PageNumberLimitExcausted) and Not(TimeLimitExchausted) )
   L = PopURLfrom(Q)
   L = normalization(L)
   If (not(Visited(L)) and isModified(L) and not(includedInQ(L)) and isAccepted(L) )
      P = downloadPage(L)
      Store a copy of P at the local repository.
      Update the index file with an entry for P.
      Visited(L) = True
      If  Parseable(P)   //   i.e. P's type is HTML
         Parse P to obtain its links. Let N be the links contained in P
         N' = those links of N that should be visited //based on accept/reject list, robots.txt, scope
         Add N' to Q // at the end for BFS, or at the beginning for DFS
         Update the index file with the eventual anchor texts
      EndIf
  EndIf
EndWhile
```

Είσοδος Ερπυστή: Το αρχείο παραμετροποίησης Crawler.conf

Το αρχείο αυτό θα περιλαμβάνει παραμέτρους που αφορούν:

- στον τρόπο ερπυσμού (**crawler-options**),
- στην καταγραφή της λειτουργίας του ερπυστή (**log-options**)
- στον επανερπυσμό (**recrawling-options**)
- στον πολυνηματισμό (**multithreading-options**)

Οι παράμετροι επεξηγούνται παρακάτω.

Crawler-Options

- **Σπόροι (seeds, starting-points):**
  Καθορίζει τους δικτυακούς τόπους από τους οποίους ο ερπυστής θα αρχίσει τη διάσχιση.
- **Φόρμουλα καθορισμού εγγράφων προς καταγραφή (Accept – List)**
  Καθορίζει, βάσει κανονικών εκφράσεων, τους αποδεκτούς (για καταγραφή) τύπους αρχείων.
- **Φόρμουλα καθορισμού εγγράφων προς απόρριψη (Reject – List)**
  Καθορίζει, βάσει κανονικών εκφράσεων, τους απορριπτέους τύπους αρχείων.
- **Μέγιστος αριθμός σελίδων προς ανάσυρση (Max Page Number)**
  Καθορίζει το μέγιστο αριθμό αρχείων που πρέπει να ανασύρει ο ερπυστής. Όταν ο ερπυστής ανασύρει αυτό το πλήθος αρχείων πρέπει να σταματά. Αν η παράμετρος αυτή έχει την τιμή 0 τότε ο ερπυστής πρέπει να συνεχίσει τη διαδικασία διάσχισης μέχρι να ανασύρει όλες τις σελίδες που μπορεί να φτάσει.
- **Scope:**
  Καθορίζει το εύρος ελευθερίας του ερπυστή σχετικά με την `αλλαγή διακομιστών'. Οι τιμές αυτής της παραμέτρου (σε φθίνουσα ως προς την ελευθερία σειρά) είναι:

  Free-spanning:

  Ο ερπυστής μπορεί να μεταβαίνει σε URL με οποιοδήποτε `domain' ή `host'.

  Domain-scope:

  Ο ερπυστής μπορεί να μεταβαίνει μόνο σε URLs που ανήκουν σε διακομιστές που ανήκουν στο ίδιο seeds-domains-set. Για παράδειγμα αν τα `seeds' είναι: {http://foo1.domain.com/foo/a.htm,
  http://foo2.foo1.domain.com/b.htm,
  http://foo3.foo1.domain.com/c/d/e/f.htm}
  τότε το `seeds-domains-set' είναι το: foo1.domain.com.

  Host-scope:

  Ο ερπυστής μπορεί να μεταβαίνει σε URLs που ανήκουν στους διακομιστές του `seeds-hosts-set'. Για παράδειγμα αν τα `seeds' είναι :
  {http://foo1.domain.com/foo/a.htm,
  http://foo2.foo1.domain.com/b.htm,
  http://foo3.foo1.domain.com/c/d/e/f.htm}
  τότε το`seeds-hosts-set' είναι το:
  {http://foo1.domain.com/,
  http://foo2.foo1.domain.com/,
  http://foo3.foo1.domain.com/}

  PathScope:

  Ο ερπυστής μπορεί να μεταβαίνει σε URLs ενός τομέα μονοπατιών που καθορίζεται από τους `σπόρους'. Φυσικά ένα διακομιστής-σπόρος που έχει URL με βάθος 1 (πχ www.sample.com/index.html) θα συμπεριληφθεί πλήρως από τη ρίζα του. Από την άλλη ένας διακομιστής του οποίου ο σπόρος είναι ο www.sample.com/path/index.html θα είναι περιορισμένος στα URL κάτω από το `/path/'.

- **Αλγόριθμος ερπυσμού διακτυακών τόπων (Traversal Algorithm)**

Τους συνδέσμους που έχει ανασύρει ο ερπυστής τους επισκέπτεται σύμφωνα με την πολιτική διάσχισης. Στη συγκεκριμένη εργασία θα χρειαστεί να υποστηρίξετε τουλάχιστον δύο πολιτικές/αλγορίθμους διάσχισης: BFS (Breadth First Search) και DFS (Depth First Search).

- **Μονοπάτι καταλόγου (repository)**

Καθορίζει το μονοπάτι καταλόγου όπου θα γίνεται η εναπόθεση των αντιγράφων που δημιουργεί ο ερπυστής από τους δικτυακούς τόπους που σαρώνει.

Log-options

- Όνομα αρχείου καταγραφής λειτουργιών: **log-file** (**π.χ** log-file = "log";)
- Βαθμός καταγραφής λειτουργιών: **log-level (π.χ** log-level = 1;).Έγκυρες τιμές είναι οι: `0, 1, 2':

`0': Μόνο τα απολύτως απαραίτητα μηνύματα τυπώνονται.

`1': Λεπτομερή μηνύματα τυπώνονται για τις ενέργειες του ερπυστή.
`2': Μηνύματα αποσφαλμάτωσης τυπώνονται για τις ενέργειες του ερπυστή.

Recrawling options

- Χρονικό διάστημα (σε ώρες) μετά το οποίο θα αρχίσει ο επόμενος ερπυσμός. Η σχετική παράμετρος είναι η **period** (**π.χ** period = 72;) και η μονάδα μέτρησης είναι η ώρα.

Η λειτουργία του Ερπυστή

Πέραν των παραμέτρων που περιγράφηκαν προηγουμένως σχετικά με τους επιτρεπτούς/απορριπτέους τύπους αρχείων, ο ερπυστής πρέπει να σέβεται το robots.txt[1] του εκάστοτε διακομιστή. Τα αρχεία `robots.txt' υπάρχουν στους διακομιστές ακριβώς για να υπαγορεύουν στους διάφορους ερπυστές που σαρώνουν το σύστημα αρχείων τους ποια αρχεία πρέπει να αποφύγει ένας ερπυστής να κατεβάσει ακόμη και αν αυτά τα αρχεία είναι κατά τα άλλα διαθέσιμα και προσπελάσιμα με ένα οποιοδήποτε φυλλομετρητή (browser).

Από κάθε αρχείο το οποίο αποδέχεται να κατεβάσει ο ερπυστής μας ενδιαφέρει να καταγράφουμε:

- Την ημερομηνία τροποποίησης και την ημερομηνία ανάσυρσης του όπως αυτές αναφέρονται στις επικεφαλίδες του Http πρωτοκόλλου. Η κύρια κλάση με την οποία θα δημιουργείτε http συνδέσεις είναι η java.net.URL.
- Την κωδικοποίηση που ακολουθεί το αρχείο
- Τους συνδέσμους που περιέχει. Ένας έτοιμος απλός parser υπάρχει στο παράδειγμα http://java.sun.com/developer/technicalArticles/ThirdParty/WebCrawler/

Ο τρόπος με τον οποίο θα καταγράφονται περιγράφεται στην επόμενη ενότητα.
Bonus: Πολυνηματική λειτουργία του ερπυστή για επιτάχυνση της όλης διαδικασίας (δημιουργία πολλών νημάτων για το κατέβασμα σελίδων).

Έξοδος Ερπυστή: Το αρχείο ευρετήριο και τα τοπικά αντίγραφα των σελίδων.

Το αποτέλεσμα της λειτουργίας του Ερπυστή θα είναι μια τοπική αποθήκη σελίδων και ένα ευρετήριο (index). Το ευρετήριο είναι ένα αρχείο με εγγραφές του τύπου:

UrlMD5checksum normalizedURL originalURL "Title" encoding type lastModifiedDate serverDate [1]
<tab> @<normalized 1st link retrieved from this page> "<anchor text for this 1st link>"
<tab> @<normalized 2nd link retrieved from this page> "<anchor text for this 2nd link>"
<tab> @<normalized 3rd link retrieved from this page> "<anchor text for this 3rd link>"

[1] http://en.wikipedia.org/wiki/robots.txt

... ...
<tab> @<normalized Nth link retrieved from this page> "<anchor text for this Nth link>"

[1]: Σε μία γραμμή

Επεξήγηση πεδίων:

- **UrlMD5CheckSum:**

  Οι `συναρτήσεις παραγωγής μηνυμάτων ταυτοποίησης' (`message digest functions') είναι συναρτήσεις κατακερματισμού. Κατηγοριοποιούνται ως `ασφαλείς και μονόδρομες'. Παίρνουν ως είσοδο δεδομένα αυθαίρετου μήκους και δίνουν ως έξοδο μία χαρακτηριστική τιμή κατακερματισμού για αυτά τα δεδομένα. Μία από τις συναρτήσεις παραγωγής μηνυμάτων ταυτοποίησης είναι η συνάρτηση MD5. Στη Java η σχετική κλάση είναι η: **java.security.MessageDigest**
- **NormalizedURL:**

  Παίρνετε την αρχική μορφή ενός URL κι έπειτα από συγκεκριμένη επεξεργασία να παράγετε την κανονικοποιημένη μορφή του. Η μέθοδος με την οποία ένα URL χάνει διάφορα περιττά χαρακτηριστικά του και φθάνει σε μία κανονικοποιημένη μορφή: [http://]<path ή IP του host>/[<κανονικοποιημένο path>] λέγεται κανονικοποίηση[2].

- **OriginalURL:**
  Το URL πριν την κανονικοποίηση.
- **"Title":**

  Ο τίτλος της σελίδας.
- **Encoding:**

  Η κωδικοποίηση της σελίδας.
- **Type:**

  Ο τύπος (type) της σελίδας (html, pdf). Προσέξτε ότι σε μερικά αρχεία υπάρχουν παραπάνω από ένας τρόποι για το προσδιορισμό του τύπου του αρχείου. Μερικές σελίδες έχουν κατάληξη `.html' άλλες `.htm' - στο πεδίο type όμως πρέπει να γραφτεί `html' και στις δύο περιπτώσεις.
- **lastModifiedDate (`last modified header field'):**

  Η τιμή αυτού του πεδίου μπορεί να βρεθεί στο πεδίο `Last-modified' που βρίσκεται στις `επικεφαλίδες' των http-απαντήσεων (http-response) του πρωτοκόλλου http. Υποδεικνύει την ημερομηνία και το χρόνο στον οποίο ο διακομιστής πιστεύει ότι τροποποιήθηκε το αρχείο στο οποίο αναφέρθηκε η αίτηση που έλαβε.
- **serverDate**
  Το σχετικό πεδίο στην επικεφαλίδα-απάντηση των διακομιστών λέγεται `Date' και αναπαριστά την ημερομηνία και τον χρόνο στον οποίο δημιουργήθηκε το μήνυμα-απάντηση του διακομιστή. Για παράδειγμα: Date: Tue, 15 Nov 1994 08:12:31 GMT
- **Λίστα αποδεκτών συνδέσμων που ανακτήθηκαν από την σελίδα:**
  Μετά από τη γραμμή που θα περιέχει όλα τα παραπάνω πεδία πρέπει να ακολουθεί μία λίστα στοιχισμένη με στηλογνώμονες (tabs) που αναφέρει όλα τα αποδεκτά links που ανακτήθηκαν από την εν λόγω σελίδα (ένα link ανά γραμμή μαζί με το anchor text του). Anchor Text είναι το κείμενο που περιέχει ένας σύνδεσμος δηλαδή: Αυτό είναι το Anchor Text

[2] http://en.wikipedia.org/wiki/URL_normalization

Η κάθε σελίδα θα έχει και τοπικό αντίγραφο το οποίο θα μπορεί να το βρει κανείς ακολουθώντας το path του URL αρχίζοντας από το βασικό κατάλογο αποθήκευσης σελίδων. Για παράδειγμα το αντίγραφο της σελίδας index.html που ανήκει στο domain `www.ics.forth.gr' θα πρέπει να βρίσκετε στο **./www.ics.forth.gr/index.html**.

Καλή εργασία!

Παράρτημα

Ενδεικτικό αρχείο παραμετροποίησης ερπυστή

```
# Configuration File for Web Crawler

<crawler-options>

# The starting points of crawler. Must be one per line ending with
# `",<newline>'
# to ease parsing.

Starting-points = {
          "http://www.csd.uoc.gr/",
          "http://www.ics.forth.gr/",
          "http://el.wikipedia.gr",
};

# accept only the following file patterns (regular expressions).
# example: "^[a-f].*\.pdf$" will download only the pdf files beginning
# with a character from a to f.
accept-list = {".*\.[hH][tT][mM][lL]?$", ".*\.[pP][hH][pP]$", ".*\.[jJ][sS][pP]$",
".*\.[aA][sS][pP]$", ".*\.[tT][xX][tT]$", ".*\.[pP][dD][fF]$", ".*\.[dD][oO][cC]$" };

# reject the following file patterns (regular expressions).
# example: "*tmp*" will not download files containing
# the string "tmp".
Reject-list = {
        ".*\.[tT][mM][pP]\.?.{0,4}$",
        "^#.*",
        ".*\?.*"
};

# the maximum number of pages to retrieve. If this number is reached the
# crawler stops. `0' or negative means no limit.
max-page-number = 1000000;

# Scope-mode that will determine the host-spanning policy:
# "free-spanning", "domain-scope", "host-scope", "path-scope"
scope = "domain-scope";

# Traversal algorithm (currently supported: bfs, dfs).
# The `depth within site' algorithm can be approximated
# by setting the algorithm here to`dfs' and setting
# `scope' to `path-scope'.
traversal-algorithm = "bfs";

# the directory where the retrieved documents will be stored.
repository = "./testbedCSDICS";
```

```
<logging>

# Log-messages target. Valid values: “log” and “screen”.
log-file = "log";

# valid values are [0-2]: 0 stands for quiet, 1 for verbose
# and 2 for debug mode
log-level = 1;

<recrawling-options>

# in hours
period = 72;
```

Πεδίο serverDate ευρετηρίου

Κατά κανόνα οι διακομιστές θα περιλαμβάνουν το πεδίο `Date' στις απαντήσεις τους με εξαίρεση τις εξής περιπτώσεις:

1. Αν η απάντηση είναι των ειδικών τύπων `Continue' (100) ή `Switching protocols' (101) η απάντηση μπορεί και να μην περιέχει το πεδίο `Date' κατά προτίμηση του διακομιστή.
2. Αν η απάντηση καταδεικνύει κάποιο λάθος όπως `Internal Server Error' (500) ή `Service Unavailable' (503) και είναι άβολο ή αδύνατο για τον διακομιστή να παράγει το πεδίο `Date'.
3. Αν ο διακομιστής δεν έχει ένα ρολόι που να μπορεί να παράγει μία αρκετά ακριβή τιμή για το πεδίο `Date' τότε η απάντηση του δε θα περιέχει καν το πεδίο.

Θεωρητικά το πεδίο `Date' φέρει τη χρονική στιγμή ακριβώς την δημιουργία του μηνύματος-απάντηση από τον διακομιστή και το χρειαζόμαστε για λόγους ελέγχου κατά την διαδικασία του recrawling.