Η Μηχανή Αναζήτησης της Ψηφιακής Βιβλιοθήκης του Θαλή

Νίκος Κουρεμένος kourem@gmail.com
Π 02161, Ιανουάριος 2006
Πανεπιστήμιο Πειραιά

Πίνακας Περιεχομένων

1. Θέμα

Να κατασκευαστεί μια σύνθετη μηχανή αναζήτησης για την ψηφιακή βιβλιοθήκη του Θαλή. Η μηχανή αναζήτησης θα ευρετηριάζει τα πεδία "λέξεις-κλειδιά" "Abstract" και "τίτλος". Να υποστηρίζονται κάποιες από τις λειτουργίες της σύνθετης αναζήτησης του Google.

2. Εισαγωγή

Η υλοποίηση χρησιμοποιεί τεχνολογίες HTML, CGI, MySQL, HTTP Server, Python. Συγκεκριμένα περιμένει την εξής δομή των αρχείων. Για παράδειγμα σε Apache η δομή των αρχείων είναι η ακόλουθη:

- httpd/html/index.html
- httpd/html/header.jpg
- httpd/html/site.css
- httpd/cgi-bin/seek.py

Ενώ το migrate_to_db.py πρέπει να καλείται από το PHP UI του Θαλή, όταν το τελευταίο ανανεώνει το αρχείο lib.xml. Συγκεκριμένα η κλήση πρέπει να γίνει με 2 παραμέτρους:

```
./migrate_to_db.py username password
```

όπου τα username και password αντιστοιχούν σε αυτά που έχουν καθοριστεί στον mysqld.

3. Απαιτήσεις

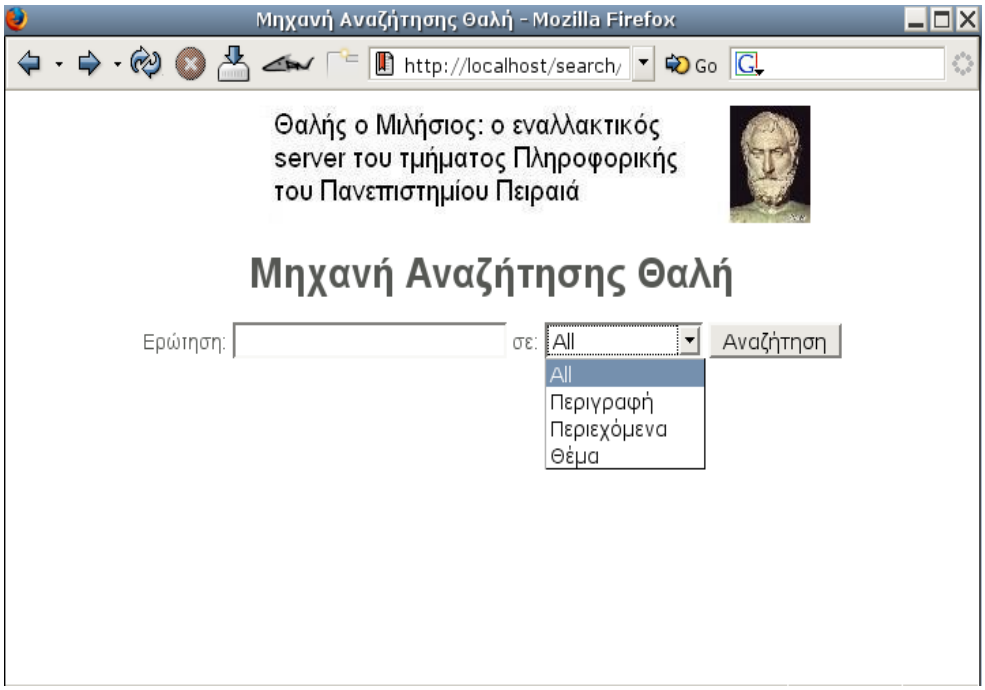
- MySQL (έκδοση 5 ή ανώτερη)
- HTTP Server με CGI-BIN στημένο (Apache, Lighttpd, thttpd, …)
- Python (έκδοση 2.3 ή ανώτερη)
- Python MySQLdb (Debian: python-mysqldb)
- Python cElementTree (Debian: python-celementtree)

4. Αρχεία

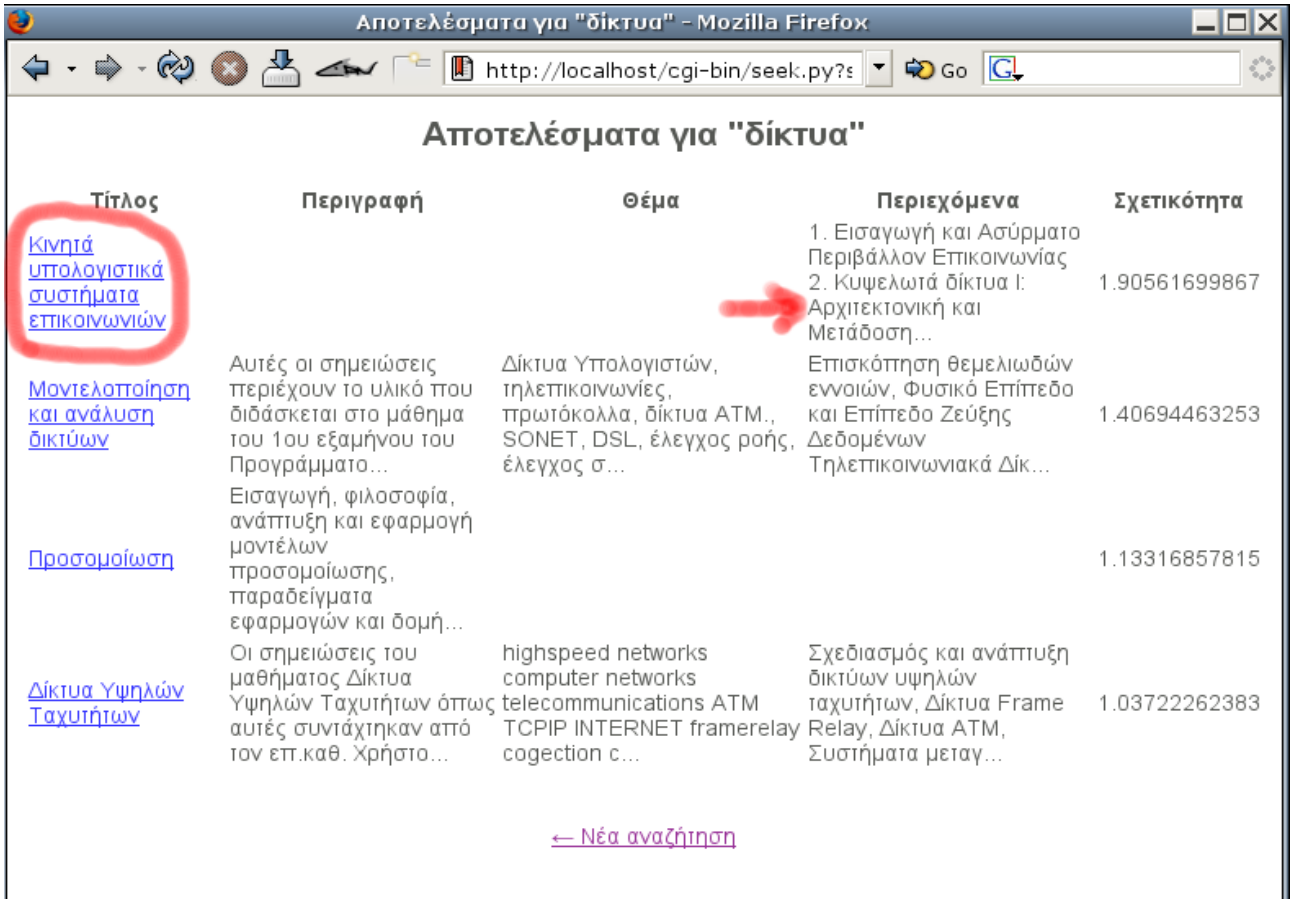
Όνομα	**Περιγραφή**
migrate_to_db.py	Διαβάζει, μετατρέπει και καταχωρεί τα δεδομένα του lib.xml σε mysql βάση με όνομα diglib και πίνακα με όνομα courses.
header.jpg (html/search/)	Το λογότυπο του Θαλή.
index.html (html/search/)	Η κύρια σελίδα της αναζήτησης, αυτή που έχει την φόρμα αναζήτησης.
site.css (html/search/)	Το css που χρησιμοποιείται για την σελίδα index.html και το seek.py.
seek.py (cgi-bin/)	Το cgi script γραμμένο σε python που λαμβάνει τις παραμέτρους από την φόρμα της index.html, ρωτάει την βάση, και μορφοποιεί τα

Όνομα	Περιγραφή
	αποτελέσματα σε html σελίδα.

5. Αναζητώντας με την Μηχανή

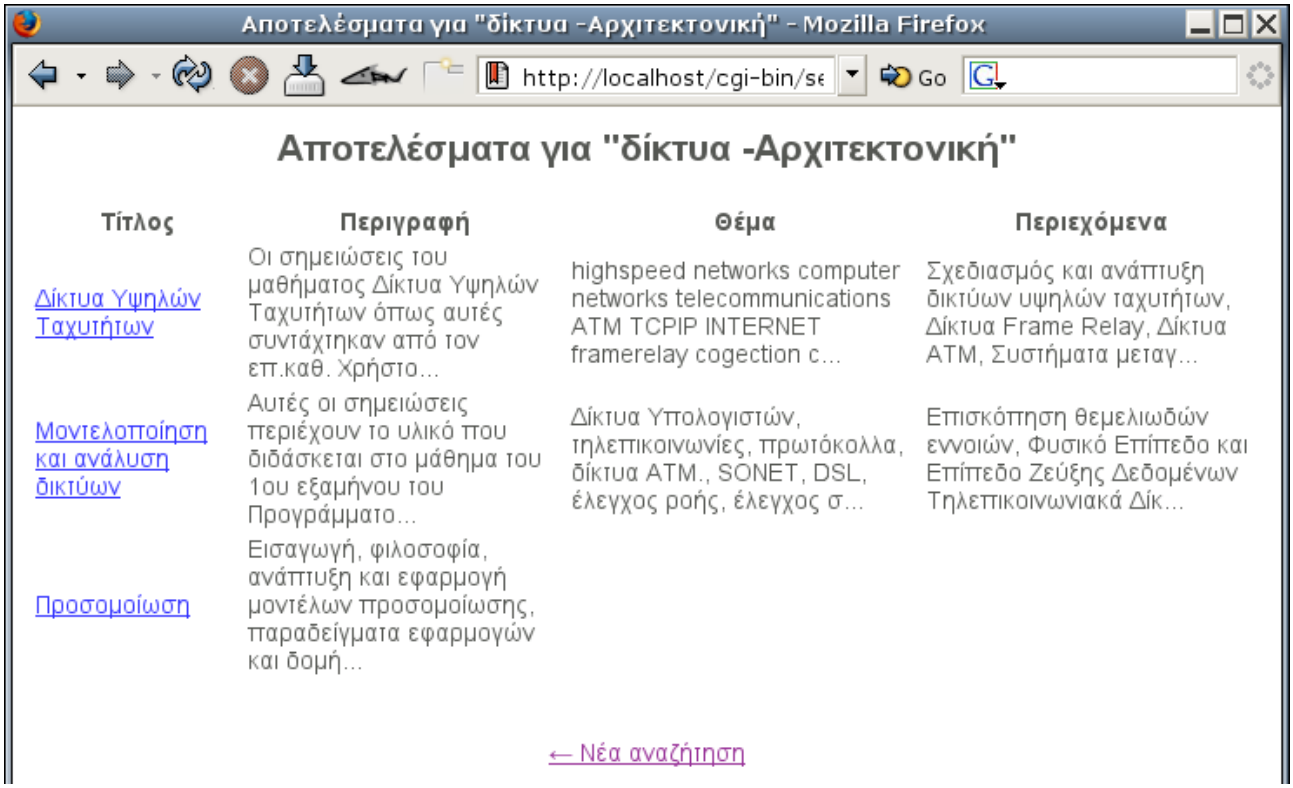


Εικόνα 1: Η βασική διεπαφή χρήστη



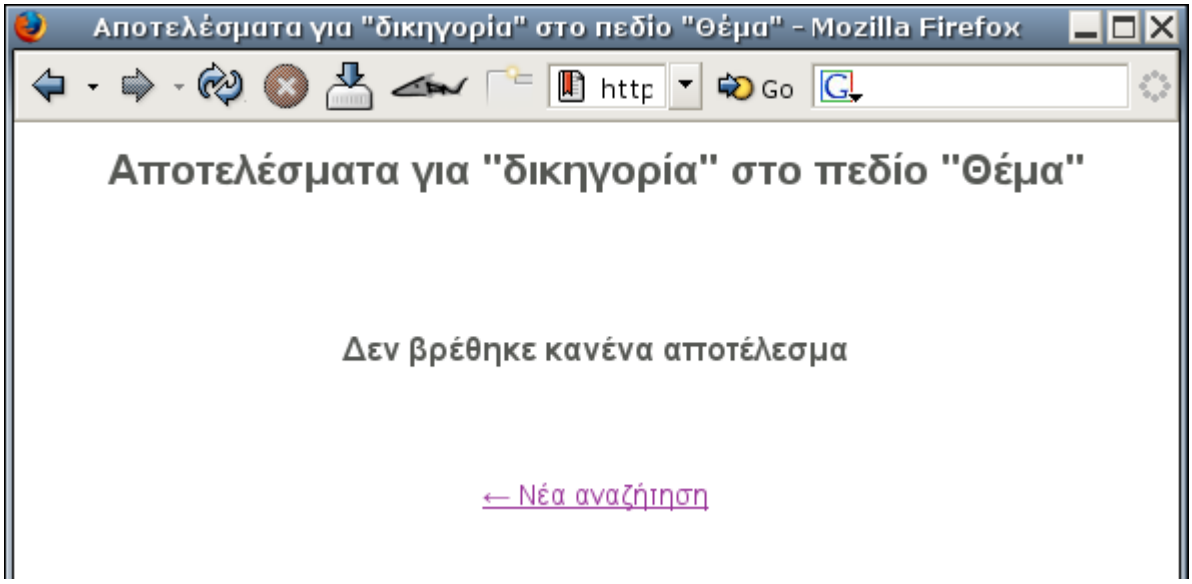
Εικόνα 2: Αναζήτηση για την λέξη δίκτυα σε όλα τα πεδία

Η MySQL δεν έχει σχετικότητα στην δυαδική αναζήτηση. Αυτή η δυνατότητα παρέχεται μόνο εφόσον χρησιμοποιηθεί επεκταμένη δυαδική αναζήτηση η οποία όμως σύμφωνα με τον Golubchik[2] "σπάει" την απόλυτη και αυστηρή ταύτιση και επιστρέφει αποτελέσματα που ταυτίζονται μερικώς. Συνεπώς στην μηχανή του Θαλή και για δυαδικές αναζητήσεις η σχετικότητα παραλείπεται όπως φαίνεται στην επόμενη εικόνα.

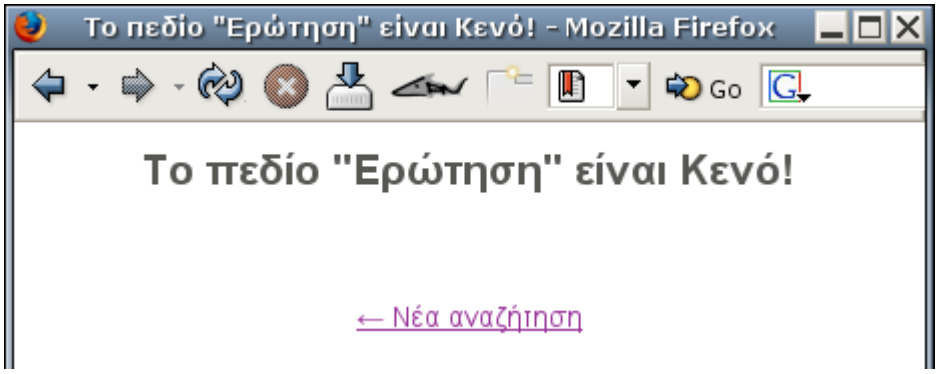


Εικόνα 3: Αναζήτηση για την λέξη δίκτυα σε όλα τα πεδία αφαιρώντας όμως όσα αποτελέσματα έχουν την λέξη Αρχιτεκτονική

Για περισσότερα πάνω στη σύνταξη των ερωτήσεων βλέπε την τεκμηρίωση της MySQL [2].



Εικόνα 4: Κανένα αποτέλεσμα για την αναζήτηση για την λέξη δικηγορία στο πεδίο Θέμα

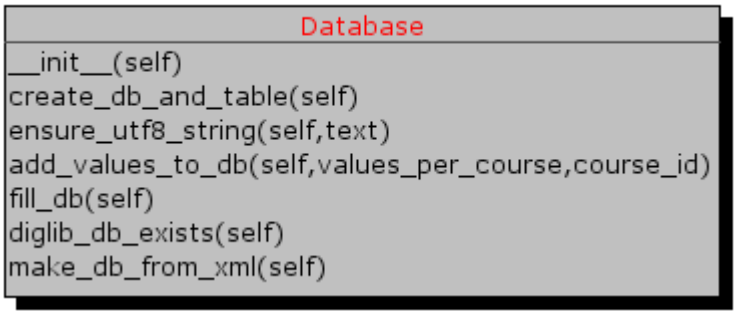


Εικόνα 5: Ειδοποίηση προς τον χρήστη για μη αποδεκτή ερώτηση

6. Σχεδιασμός Κώδικα

6.1 Σύντομη Περιγραφή

Το αρχείο **migrate_to_db.py** έχει την τάξη *Database* με το παρακάτω UML:



Η βασική μέθοδος είναι η *make_db_from_xml* η οποία καλεί την *diglib_db_exists* και αν γυρίσει True αφαιρεί την βάση. Ανεξάρτητα από την τιμή επιστροφής της *diglib_db_exists* γίνεται κλήση της *create_db_and_table*. Η μέθοδος *create_db_and_table* συνδέεται στον mysqld και δημιουργεί την βάση *diglib* τον πίνακα courses και ορίζει την κωδικοποίηση σε Unicode (UTF-8). Στην συνέχεια καλείται η *fill_db*.

Η *fill_db* αναλύει (parse) το lib.xml αρχείο και για αφού συγκεντρώσει τις τιμές που απετούνται για μια εγγραφή καλεί την *add_values_to_db* η οποία και κάνει το EXECUTE. Όταν έχει διαβαστεί όλο το lib.xml αρχείο και έχουν γίνει όλα τα EXECUTE τότε γίνεται ένα COMMIT με το οποίο καταχωρούνται στην βάση τα δεδομένα. Αυτό γίνεται μια φορά και στο τέλος για λόγους αποδοτικότητας και ταχύτητας της εφαρμογής.

Τέλος η *ensure_utf8_string* μετατρέπει το κείμενο που της δίνουμε σε utf8 αν αυτό χρειάζεται, αφού όπως προαναφέρθηκε όλα τα δεδομένα πρέπει να είναι utf8.

Το αρχείο **seek.py** είναι το αρχείο που κάνει λαμβάνει τα δεδομένα από την φόρμα αναζήτησης, κάνει την ερώτηση στη βάση και μορφοποιεί το ανάλογο αποτέλεσμα. Η βασική συνάρτηση που καλεί με τις ανάλογες παραμέτρους τις ανάλογες συναρτήσεις αναλύεται παρακάτω. Όταν τυπώνουμε (print) το stdout ανακατευθύνεται από τον εξυπηρετητή HTTP στον πελάτη HTTP (περισσότερα στα σχόλια του ίδιου του αρχείου):

```
print 'Content-type: text/html; charset=UTF-8\n'
form = cgi.FieldStorage()
```

```
query = form.getlist('search_entry')
if len(query) == 0:
      text = 'Το πεδίο "Ερώτηση" είναι Κενό!'
      print_inital_and_title_and_text(text)
      print_go_back()
      print '</body></html>'
      return
else:
      query = query[0]
kind = form.getlist('kind_of_search')[0]

query = decode_html(query)
if query.find('+') != -1 or query.find('-') != -1 or\
      query.find('*') != -1:
      # if we have + or - or *
      # search in boolean mode
      results = query_db(kind, query, mode = 'boolean')
      show_results(results, kind, query, mode = 'boolean')
else:
      results = query_db(kind, query, mode = 'normal')
      show_results(results, kind, query, mode = 'normal')
```

6.2. Δημιουργία Πίνακα courses

```
CREATE TABLE courses(
        row_id INTEGER PRIMARY KEY AUTO_INCREMENT UNIQUE,
        course_id INTEGER UNIQUE,
        title TEXT,
        creator TEXT,
        educationLevel TEXT,
        description TEXT,
        tableOfContents TEXT,
        created INTEGER,
        modified INTEGER,
        subject TEXT,
        publisher TEXT,
        identifier TEXT,
        FULLTEXT (title, description, tableOfContents, subject),
        FULLTEXT (title, description),
        FULLTEXT (title, tableOfContents),
        FULLTEXT (title, subject)
)
DEFAULT CHARACTER SET utf8 COLLATE utf8_unicode_ci;
```

Τα ονόματα των στηλών του πίνακα είναι ίδια με αυτά του lib.xml αρχείου.

6.3. Μια FULLTEXT Ερώτηση

Παράδειγμα ερώτησης με πρώτο αποτέλεσμα το πιο σχετικό στην αναζήτηση για *xml* σε όλα τα πεδία:.

```
SELECT *, MATCH (title, description, tableOfContents, subject)
AGAINST ('xml') AS score
FROM courses WHERE
MATCH (title, description, tableOfContents, subject)
AGAINST ('xml')
```

7. Αναφορές

[1] Sergei Golubchik. Int. PHP Conference: Fulltext Search, Frankfurt, 2003 προσβάσιμη στην διεύθυνση http://www.php-kongress.de/2003/slides/database_track/golubchik_mysql_fulltext_search_2003.pdf.

[2] MySQL AB, MySQL 5.0 Reference Manual: Full-Text Search Functions προσβάσιμη στην διεύθυνση http://dev.mysql.com/doc/mysql/en/Fulltext_Search.html.