Τμήμα Μηχανικών Η/Υ και Πληροφορικής

Πανεπιστήμιο Πατρών

ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ ΙΙ

ΕΡΓΑΣΙΑ ΕΞΑΜΗΝΟΥ

Εαρινό Εξάμηνο 2013

Διαχείριση Δεδομένων Χρονοσειρών

Ημερομηνία Παράδοσης: 3-07-2013

Βαρύτητα: 30%

Περιγραφή

Στόχος της εργασίας είναι η δημιουργία μιας εφαρμογής διαχείρισης δεδομένων χρονοσειρών. Η εφαρμογή θα δίνει την δυνατότητα σε χρήστες να αποθηκεύουν χρονοσειρές και να πραγματοποιούν διάφορα ερωτήματα για την ανάκτησή τους. Βασικό χαρακτηριστικό της εφαρμογής είναι η δυνατότητα εκτέλεσης «ερωτημάτων βάσει περιεχομένου» (query by content).

Τα βασικά τμήματα της εργασία είναι η δημιουργία **α)** μια βάσης δεδομένων στην οποία θα αποθηκεύονται δεδομένα χρονοσειρών και **β)** μιας εφαρμογής που θα διαχειρίζεται μέσω γραφικού περιβάλλοντος τις επιλογές των χρηστών.

Σύνολο Δεδομένων

Χαρακτηριστικά παραδείγματα δεδομένων χρονοσειρών είναι:

- Ηλεκτροεγκεφαλογράφημα(EEG)
- fMRI – functional Magnetic Resonance Imaging
- Ήχος (μουσική, ομιλία κλπ)

Η κάθε ομάδα έχει την δυνατότητα επιλογής του συνόλου δεδομένων που θα χρησιμοποιήσει. Εκτός από τις χρονοσειρές, το σύνολο δεδομένων που θα επιλεχθεί θα πρέπει να έχει και κάποια μετα-δεδομένα που προσδιορίζουν τις χρονοσειρές. Για παράδειγμα, τα αρχεία μουσικής έχουν τίτλο τραγουδιού, όνομα καλλιτέχνη, album, κ.λ.π.. Στην εικόνα 1 φαίνεται η λογική δομή των δεδομένων ηλεκτροεγκεφαλογραφήματος που μπορούν να χρησιμοποιηθούν στην άσκηση.

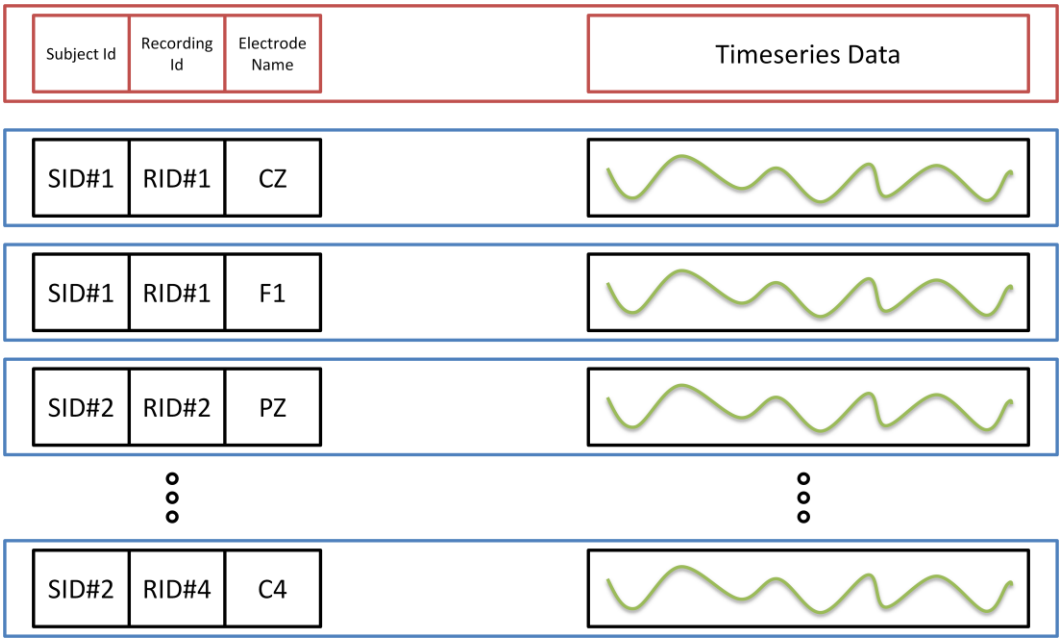

Figure 1 Logical Structure of the EEG data

Αντίστοιχη με την παραπάνω λογική δομή θα πρέπει να ισχύει για οποιοδήποτε σύνολο δεδομένων επιλέξετε.

Επιλογή του συνόλου δεδομένων

Το σύνολο δεδομένων που θα επιλέξετε θα καθορίσει σε μεγάλο βαθμό την υλοποίηση της άσκησης. Αν επιλέξετε κάποια από τις προτεινόμενες κατηγορίες, μπορείτε να βρείτε δεδομένα από τις παρακάτω πηγές:

- EEG
  - http://physionet.org/physiobank/database/
- Ήχος
  - http://grh.mur.at/sites/default/files/mir_datasets_0.html#music-information-retrieval-datasets

Για δεδομένα EEG, fMRI μπορείτε να επικοινωνήσετε με τους υπεύθυνους του εργαστηρίου. Επίσης, οι παραπάνω πηγές δεν είναι δεσμευτικές. Μπορείτε να χρησιμοποιήσετε το σύνολο δεδομένων που προτιμάτε.

Εξαγωγή Χαρακτηριστικών

Έχοντας επιλέξει το σύνολο δεδομένων, το επόμενο βήμα είναι η επιλογή της μεθόδου εξαγωγής χαρακτηριστικών που θα χρησιμοποιήσετε στην εφαρμογή σας. Το συγκεκριμένο βήμα προτείνουμε να το πραγματοποιήσετε σε Matlab, Octave ή Python.

Κάθε χρονοσειρά θα πρέπει να χωριστεί σε τμήματα και για κάθε ένα τμήμα θα γίνεται ο υπολογισμός των χαρακτηριστικών. Το μέγεθος του τμήματος καθορίζεται από τα δεδομένα που θα χρησιμοποιηθούν και από την μέθοδο εξαγωγής των χαρακτηριστικών που θα επιλεχθεί.

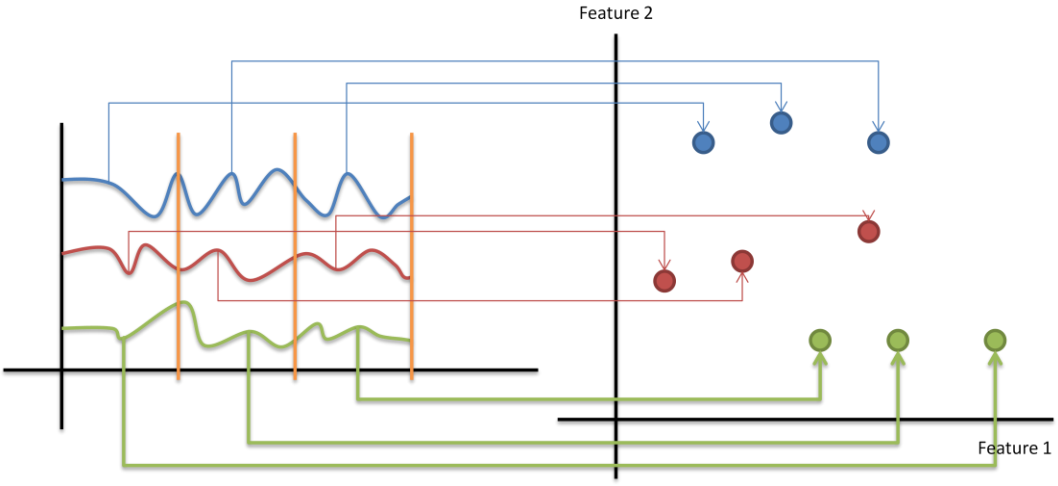

Figure 2 Feature Extraction Example (2 Features)

Στο συγκεκριμένο βήμα θα πρέπει να γίνει η επιλογή των παραμέτρων της εφαρμογής που θα δημιουργηθεί καθώς και η αξιολόγηση της μεθόδου εξαγωγής χαρακτηριστικών που θα επιλεχθεί.

Δεικτοδότηση Δεδομένων

Για την αναζήτηση βάση περιεχομένου (query by content) θα πρέπει να χρησιμοποιήσετε μια μέθοδο χωρικής προσπέλασης - Spatial Access Method (SAM), όπως παρουσιάστηκε στο μάθημα (προτείνεται το R-Tree). Μπορείτε να χρησιμοποιήσετε κάποια SAM σε περίπτωση που αυτή παρέχεται από το ΣΔΒΔ που θα χρησιμοποιηθεί. Σε αντίθετη περίπτωση θα πρέπει είτε να υλοποιήσετε εσείς τη δομή είτε να χρησιμοποιήσετε κάποια έτοιμη υλοποίηση που θα βρείτε στο διαδίκτυο.

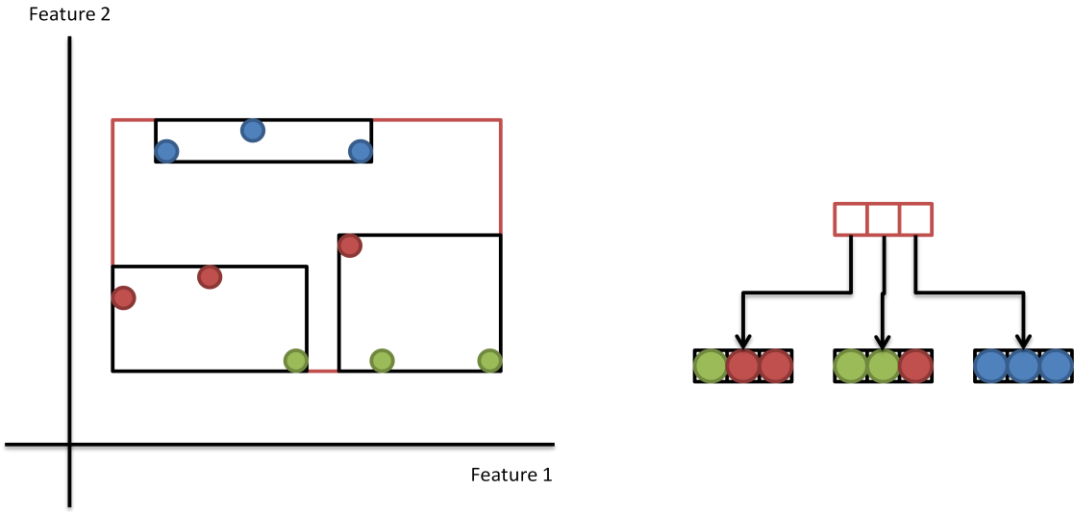

Figure 3 Spatial Indexing of Features

Η συνάρτηση απόστασης στον χώρο των χαρακτηριστικών θα είναι η Ευκλείδεια.

Αποθήκευση

Για την αποθήκευση των μετα-δεδομένων θα πρέπει να χρησιμοποιηθεί ένα DBMS (προτείνεται η PostgreSQL). Για την αποθήκευση των αρχείων που θα περιέχουν τα δεδομένα των χρονοσειρών μπορείτε να χρησιμοποιήσετε το ίδιο DBMS αποθηκεύοντας είτε το path στο αρχείο είτε binary large objects - blobs.

Υλοποίηση

Η υλοποίηση μπορεί να πραγματοποιηθεί σε οποιαδήποτε γλώσσα προγραμματισμού επιθυμείτε. Η εφαρμογή που θα υλοποιήσετε θα πρέπει να παρέχει με γραφικό περιβάλλον στον χρήστη τις παρακάτω δυνατότητες:

- Εγγραφή/Σύνδεση στο σύστημα
- Πραγματοποίηση ερωτήματος με βάση κάποιο πεδίο των δεδομένων. Το αποτέλεσμα αυτού του ερωτήματος θα πρέπει να είναι μία λίστα με τις χρονοσειρές που ικανοποιούν το ερώτημα καθώς και ένας σύνδεσμος στο πραγματικό αρχείο δεδομένων.
- **Ερώτημα με βάση το περιεχόμενο της χρονοσειράς**. Το συγκεκριμένο ερώτημα θα δέχεται ως είσοδο 2 παραμέτρους:
    - Ένα αρχείο που θα περιέχει την χρονοσειρά – ερώτημα
    - Τον αριθμό των k πιο «όμοιων» χρονοσειρών που επιθυμεί ο χρήστης.

  Το αποτέλεσμα αυτού του ερωτήματος θα πρέπει να είναι οι k χρονοσειρές και τα meta-data τους. Η παρουσίαση του αποτελέσματος θα πρέπει να γίνει με γραφική αναπαράσταση.

Παραδοτέα

Οι ομάδες θα πρέπει να παραδώσουν:

- Μια σύντομη αναφορά με την περιγραφή των λειτουργιών που υλοποιήθηκαν και τις μεθόδους που χρησιμοποιήθηκαν
- Τον κώδικα της εφαρμογής

Επίσης, θα πραγματοποιηθεί εξέταση πάνω στην εργασία σε ημερομηνία που θα οριστεί μετά τη γραπτή εξέταση του μαθήματος.

Χρήσιμοι Σύνδεσμοι

- http://www.dblab.upatras.gr/download/courses/db2/Slides/5_MultimediaIndexing.pdf
- http://www.dblab.upatras.gr/download/courses/db2/2012/2_Multimedia_Indexing.pdf
- http://www.postgresql.org/
- http://labrosa.ee.columbia.edu/millionsong/

Για απορίες σχετικά με την εργασία:

korbesis@ceid.upatras.gr