ΘΕΜΑΤΑ ΔΙΠΛΩΜΑΤΙΚΩΝ ΕΡΓΑΣΙΩΝ

Εργ. Συστημάτων Βάσεων Γνώσεων & Δεδομένων

ΑΝΑΛΥΣΗ ΔΙΑΔΙΚΤΥΑΚΩΝ ΔΕΔΟΜΕΝΩΝ ΠΟΛΥ ΜΕΓΑΛΗΣ ΚΛΙΜΑΚΑΣ ΜΕ ΧΡΗΣΗ ΤΕΧΝΟΛΟΓΙΩΝ ΥΠΟΛΟΓΙΣΤΙΚΟΥ ΝΕΦΟΥΣ (BIG DATA CLOUD ANALYTICS)

ΠΛΗΡΟΦΟΡΙΕΣ: Κόνιαρης Μάριος, 210 772 1402, mkoniari –at- dblab –dot- ntua -dot- gr
Κόλιας Βασίλειος, 210 772 2538, vkolias -at- medialab -dot- ntua -dot- gr

ΠΕΡΙΛΗΨΗ: Σκοπός της εργασίας είναι η μελέτη, δομική ανάλυση και η εξαγωγή στατιστικών για το www, με βάση τη συλλογή Common Crawl.

ΑΤΟΜΑ: 2 [Μπορούν να δοθούν και 2 ξεχωριστές εργασίες]
ΠΛΑΤΦΟΡΜΑ ΕΡΓΑΣΙΑΣ: Hadoop, Map Reduce, Java, NOSQL databases.

ΣΥΝΤΟΜΗ ΠΕΡΙΓΡΑΦΗ: Big Data καλούνται τα σύνολα δεδομένων που είναι πολύ μεγάλα σε μέγεθος και είναι δύσκολο να τα διαχειριστούμε με παραδοσιακές μεθόδους. Το Common Crawl[1] διατηρεί και διαθέτει στο κοινό -μέσω του Amazon EC2/S3- ένα ψηφιακό αποθετήριο δεδομένων του παγκόσμιου ιστού. Το αποθετήριο αυτό περιλαμβάνει περίπου 4 με 5 δισεκατομμύρια σελίδες (μόνο για το 2012), συνολικού μεγέθους άνω των 100 TB. Το νέφος υπολογιστών (cloud), περιγράφεται ως ένα σύνολο υπολογιστικών συστημάτων, συνδεδεμένα μεταξύ τους με τέτοιο τρόπο ώστε να προσφέρουν δυναμικές και επεκτάσιμες υποδομές, για την εκτέλεση απαιτητικών υπολογισμών και την αποθήκευση μεγάλου όγκου δεδομένων. Το αντικείμενο της διπλωματικής είναι η δομική ανάλυση του Common Crawl και η εξαγωγή στατιστικών για τον παγκόσμιο ιστό και τις ιδιότητές του. Αποσκοπεί στο να επεκτείνει προϋπάρχουσα εργασία και μελέτη στο πρόβλημα αυτό [1,2] και θα πραγματοποιηθεί σε συνεργασία και με το Εργαστήριο Τεχνολογίας Πολυμέσων του ΕΜΠ και το Εργαστήριο Τηλεπικοινωνιών, Δικτύων και Ενοποιημένων Υπηρεσιών του Πανεπιστημίου Θεσσαλίας. Η αναγκαία υποδομή θα καλυφθεί από τα cloud services okeanos της ΕΔΕΤ[2]. Έχοντας πρότερη γνώση σε περιβάλλον cloudera enterprise:

- Θα σχεδιαστεί και θα υλοποιηθεί η ροή εργασίας σε map reduce για την εξαγωγή στατιστικών του αποθετηρίου (π.χ. σύνολο σελίδων ανά γλώσσα, τύπο τεχνολογίας, δομή in/out-degree κ.α.),
- θα μελετηθεί η σχετική βιβλιογραφία και θα συζητηθούν προσεγγίσεις/ ιδέες για την αυτόματη κατηγοριοποίηση (classification) ιστοσελίδων,
- θα υλοποιηθούν αλγόριθμοι και κριτήρια κατηγοριοποίησης ιστοσελίδων και θα αξιολογηθεί η αποτελεσματικότητά τους.

Οι ενδιαφερόμενοι θα πρέπει να έχουν καλή γνώση σε Java και τεχνολογίες cloud (π.χ. Hadoop[3], Map Reduce, HDFS, Hive), καθώς και εμπειρία σε τεχνικές Information Retrieval.

ΣΧΕΤΙΚΟ ΥΛΙΚΟ

[1]. Statistics of the Common Crawl Corpus 2012, Technical Report, https://docs.google.com/file/d/1_9698uglerxB9nAglvaHkEgU-iZNm1TvVGuCW7245-WGvZq47teNpb_uL5N9/edit

[2]. V. Kolias, I. Anagnostopoulos, E. Kayafas, Exploratory Analysis of a Terabyte Scale Web Corpus, http://arxiv.org/abs/1409.5443

[1] http://commoncrawl.org/

[2] https://okeanos.grnet.gr/home/

[3] http://en.wikipedia.org/wiki/Apache_Hadoop