

ΘΕΜΑΤΑ ΔΙΠΛΩΜΑΤΙΚΩΝ ΕΡΓΑΣΙΩΝ
Εργαστήριο Συστημάτων Βάσεων Γνώσεων & Δεδομένων

ΕΠΟΠΤΕΙΑ ΓΕΩΓΡΑΦΙΚΗΣ ΕΞΑΠΛΩΣΗΣ ΣΥΖΗΤΗΣΕΩΝ ΣΕ ΚΟΙΝΩΝΙΚΑ ΔΙΚΤΥΑ

ΠΛΗΡΟΦΟΡΙΕΣ: Κώστας Πατρούμπας, 210 772 1446, kpatro@dblab.ece.ntua.gr

ΠΕΡΙΛΗΨΗ: Στην εργασία θα σχεδιασθεί και θα υλοποιηθεί μεθοδολογία παρακολούθησης της χωρο-χρονικής εξέλιξης μηνυμάτων σε κοινωνικά δίκτυα, όπου σχηματίζονται ευάριθμες, αλλά εφήμερες «κοινοότητες» χρηστών που σχολιάζουν το ίδιο –γεωγραφικά προσδιορισμένο– θέμα.

ΑΤΟΜΑ: 1

ΠΕΡΙΒΑΛΛΟΝ ΕΡΓΑΣΙΑΣ: C++/ JAVA

ΣΥΝΤΟΜΗ ΠΕΡΙΓΡΑΦΗ: Τα μηνύματα που δημοσιεύουν οι χρήστες *κοινωνικών δικτύων* (*social networks*) έχουν συνήθως επίκαιρο χαρακτήρα, λ.χ. σχολιάζουν ένα γεγονός που μόλις συνέβη, πυροδοτώντας συχνά μία ακολουθία σχολίων για το ίδιο θέμα (*trending topic*). Ωστόσο, πολλές φορές μία «*συζήτηση*» μ' αυτήν την θεματολογία έχει και *χωρική πτυχή*, όταν τα μηνύματα συνοδεύονται από γεωγραφικό προσδιορισμό, όπως η τρέχουσα θέση του σχολιαστή (*geo-tagged tweets*) ή αναφορά μιας τοποθεσίας στο κείμενο (λ.χ., #Syntagma), η οποία προφανώς συνδέεται με το περιεχόμενο του μηνύματος (π.χ. διαδήλωση στην πλατεία Συντάγματος). Έχει λοιπόν ενδιαφέρον να μελετηθεί *πώς εξελίσσονται τέτοιες συζητήσεις από χωροχρονική σκοπιά*, παρατηρώντας την μεταβαλλόμενη έκταση και την χωρική διακύμανση των σχετικών μηνυμάτων που καταφθάνουν ως *ρεύμα δεδομένων* (*data stream*) με την πάροδο του χρόνου. Για παράδειγμα, τα μηνύματα σχετικά με μία πυρκαγιά αρχικά προέρχονται από παρατηρητές κοντά στην αρχική εστία, αλλά σταδιακά μπορεί να καλύψουν ευρύτερη περιοχή (ίσως με διαφορετική συχνότητα κατά τόπους), και τελικά εκλείπουν όταν η πυρκαγιά τεθεί υπό έλεγχο.

Σύμφωνα με πρόσφατες επιστημονικές διερευνήσεις, η διασπορά και η χωρική έκταση συσχετιζόμενων μηνυμάτων μπορεί ν' αποκαλύψει την ένταση του φαινομένου και την πιθανή επίδρασή του στο κοινωνικό περιβάλλον, προκειμένου να βελτιωθεί η ενημερότητα των χρηστών σε γεγονότα και καταστάσεις με ισχυρό τοπικό αντίκτυπο. Γι' αυτόν τον σκοπό, θα επιχειρηθεί ο σχεδιασμός μιας μεθόδου που λαμβάνει υπ' όψιν το περιεχόμενο των κειμένων (πιθανόν μέσω καταλόγου από χαρακτηριστικές λέξεις – *keywords* ή *hashtags*), την χρονική τους διάρκεια, καθώς και την χωρική τους κατανομή, με στόχο την αποτελεσματική *συνάθροιση* (*aggregation*) και την ανίχνευση της εξέλιξής τους.

Τα στάδια εκπόνησης της διπλωματικής εργασίας προβλέπονται ως εξής:

- Βιβλιογραφική επισκόπηση μεθόδων διαχείρισης χωρο-χρονικών ρευμάτων δεδομένων και επεξεργασίας χωρο-κειμενικών ερωτημάτων (*spatio-textual queries*).
- Υλοποίηση ενός *ευρετηρίου* (*index*) τήρησης των δεδομένων στην κύρια μνήμη, το οποίο θα μπορεί να συνδυάζει την χωρική εγγύτητα με την συνάφεια περιεχομένου των μηνυμάτων. Αυτό θα μπορούσε να γίνει με χρήση τετραδικού δένδρου (*quad-tree*) ή καννάβου (*grid*), σε κάθε κόμβο των οποίων θα αντιστοιχούνται οι πρόσφατες λέξεις-κλειδιά με αξιοσημείωτη συχνότητα.
- Με βάση το ευρετήριο, θα είναι εφικτή η αναγνώριση των δημοφιλών (*trending*) συζητήσεων και θα μπορεί να εκτιμηθεί η χωρική τους εξάπλωση. Όλη η επεξεργασία πρέπει να διεξάγεται *online*, οπότε ο αλγόριθμος θα πρέπει να επιστρέφει περιοχές με κυμαινόμενη έκταση και διαβάθμιση εντός ενός *χρονικού παραθύρου* (λ.χ. κατά την τελευταία ώρα).
- Αξιολόγηση πειραματικών αποτελεσμάτων εκτελώντας προσομοιώσεις με δοκιμαστικά δεδομένα και μετρώντας χρόνους απόκρισης, απαιτήσεις σε μνήμη, ακρίβεια απαντήσεων κ.ά.