Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων
Εργαστηριο Υπολογιστικων Συστηματων

Διπλωματική Εργασία 2012

Χωροχρονικός Συσχετισμός Πληροφοριών Κοινωνικών Δικτύων

Εισαγωγή

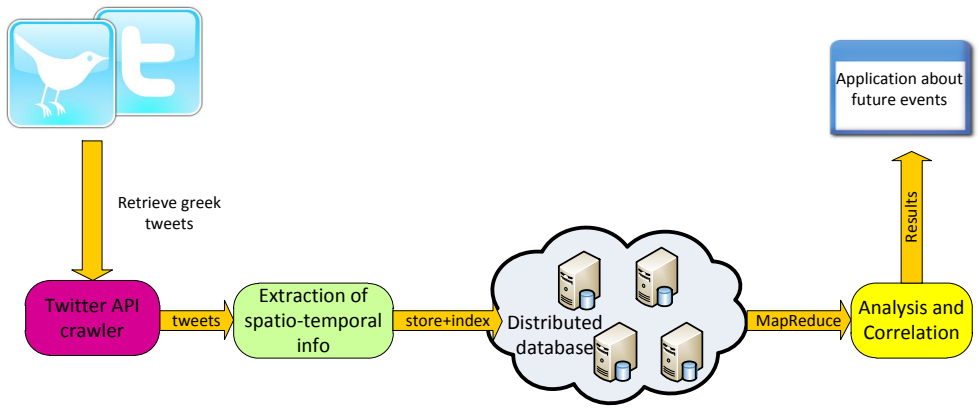
Τα κοινωνικά δίκτυα (π.χ. Facebook, MySpace, Twitter κλπ.) είναι υπηρεσίες που παρέχουν στους χρήστες τους τη δυνατότητα να συνδέονται με φίλους τους και να μοιράζονται με αυτούς πληροφορίες. Το Twitter [5], ένα ιδιαίτερο κοινωνικό δίκτυο που αυτοαποκαλείται υπηρεσία microblogging, επιτρέποντας στους χρήστες του να δημοσιοποιούν μηνύματα με μέγιστο μέγεθος 140 χαρακτήρων (tweets), αλλά και να ακολουθούν (follow) τα tweets των χρηστών με τους οποίους είναι συνδεδεμένοι.

Με την αύξηση της δημοφιλίας του Twitter, τόσο η επιστημονική κοινότητα όσο και ο επιχειρηματικός κόσμος αντιλαμβάνεται τη σημασία που έχει η ανάλυση των πληροφοριών που παράγει για την διαπίστωση νέων τάσεων και κυρίως για την πρόβλεψη γεγονότων όπως για παράδειγμα την κυκλοφοριακή συμφόρηση σε σημεία της πόλης λόγω κάποιας εκδήλωσης. Για να επιτευχθεί κάτι τέτοιο, είναι απαραίτητη η συλλογή χωροχρονικής πληροφορίας και η ανάλυσή της για τη δημιουργία συσχετίσεων. Αν για παράδειγμα μεγάλος αριθμός από tweets αναφέρονται στην 17η Μαΐου και το Παναθηναϊκό Στάδιο, τότε πιθανόν να γίνεται κάποια εκδήλωση (η τελετή παράδοσης της Ολυμπιακής φλόγας) και να απαιτούνται ειδικές κυκλοφοριακές ρυθμίσεις.

Σχήμα 1: Παράδειγμα tweet με χωροχρονική πληροφορία

Με τον αριθμό των χρηστών του Twitter να ξεπερνά τα 140 εκατομμύρια και να αυξάνεται συνεχώς, είναι εμφανές ότι υπάρχει ένας τεράστιος όγκος δεδομένων προς ανά-

λυση. Στην παρούσα διπλωματική θα ασχοληθούμε με την μεγάλης κλίμακας ανάλυση δημοσιευμένων tweets για την εξαγωγή συσχετίσεων στο χώρο και στο χρόνο, εξερευνώντας τεχνικές από την περιοχή των κατανεμημένων συστημάτων.



Σχήμα 2: Αρχιτεκτονική του συστήματος

Σκοπός
Σκοπός της διπλωματικής είναι η ανάπτυξη ενός συστήματος συλλογής, αποθήκευσης και ανάλυσης χωροχρονικών δεδομένων από tweets ώστε να εξάγονται συσχετίσεις χώρου-χρόνου και πληροφορίες για μελλοντικά γεγονότα. Γι' αυτόν το σκοπό θα γίνει χρήση τεχνικών από την περιοχή των Κατανεμημένων Συστημάτων (Cloud/P2P, π.χ. [1–3]). Η πορεία που θα ακολουθηθεί είναι η εξής:

1. Ανάπτυξη crawler για την συλλογή μεγάλου αριθμού πρόσφατων tweets, γραμμένων στα ελληνικά, κάνοντας χρήση του Twitter API [4], κατά προτίμηση σε Java.

2. Συλλογή των tweets και εξαγωγή των χωροχρονικών πληροφοριών σύμφωνα με τα οποία θα τα δεικτοδοτήσουμε.

3. Ανάπτυξη σε Java κατανεμημένου συστήματος ανάλυσης των πληροφοριών αυτών με σκοπό τη συσχέτισή τους και την εξαγωγή προβλέψεων.

4. Αξιολόγηση του συστήματος μέσω μελέτης περιπτώσεων (case studies).

Επικοινωνία:
Κατερίνα Δόκα, katerina@cslab.ece.ntua.gr
Δημήτρης Τσουμάκος, dtsouma@cslab.ece.ntua.gr

Βιβλιογραφία

[1] The Apache HBase Project. `http://hbase.apache.org/`. 2

[2] J. Dean and S. Ghemawat. Mapreduce: Simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008. 2

[3] Ashish Thusoo, Joydeep Sen Sarma, Namit Jain, Zheng Shao, Prasad Chakka, Suresh Anthony, Hao Liu, Pete Wyckoff, and Raghotham Murthy. Hive - A Warehousing Solution Over a Map-Reduce Framework. *PVLDB*, 2(2):1626–1629, 2009. 2

[4] Twitter API wiki. http://apiwiki.twitter.com/. 2

[5] Twitter Webpage. http://twitter.com/. 1