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

Διπλωματική Εργασία 2010-2011

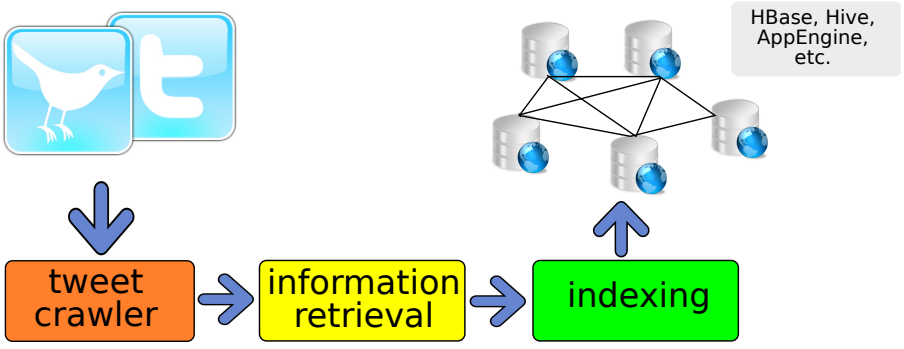
Κατανεμημένη Αποθήκευση και Δεικτοδότηση Πληροφοριών Κοινωνικών Δικτύων

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, τον προσδιορισμό προτύπων συμπεριφοράς και τη σφυγμομέτρηση του κοινού. Για παράδειγμα η παρακολούθηση και η ανάλυση του περιεχομένου των tweets καθώς και των συνηθειών των χρηστών (sentiment and behavioral analysis) μπορεί να βοηθήσει μια εταιρία να αξιολογήσει την αποδοχή ενός προϊόντος της από το κοινό ή να διαμορφώσει κατάλληλα τη στρατηγική marketing για να προσελκύσει αγοραστές.

Με τον αριθμό των χρηστών του Twitter να ξεπερνά τα 75 εκατομμύρια και να αυξάνεται συνεχώς, είναι εμφανές ότι υπάρχει ένας τεράστιος όγκος δεδομένων προς ανάλυση. Στην παρούσα διπλωματική θα ασχοληθούμε με την αποδοτική δεικτοδότηση και αποθήκευση πληροφοριών σχετικών με δημοσιευμένα tweets εξερευνώντας τεχνικές από την περιοχή των κατανεμημένων συστημάτων.



Σχήμα 1: Αρχιτεκτονική του συστήματος

Σκοπός
Σκοπός της διπλωματικής είναι η ανάπτυξη ενός συστήματος αποθήκευσης και δεικτοδότησης ενός μεγάλου αριθμού από tweets με βάση κάποια σημαντικά χαρακτηριστικά τους (όπως ο χρόνος ανάρτησης, ο συγγραφέας, οι λέξεις κλειδιά που περιέχουν κλπ.) αλλά και η διατήρηση συναθροισμένων αριθμητικών στοιχείων που τα αφορούν (π.χ. ο αριθμός των followers κλπ.) [2] ώστε να διευκολύνεται η αναζήτησή και η ανάλυσή τους. 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 κάνοντας χρήση του Twitter API [4], κατά προτίμηση σε Java.

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

3. Ανάπτυξη σε Java κατανεμημένου συστήματος αποθήκευσης των πληροφοριών αυτών με σκοπό την αποδοτική αναζήτησή τους.

4. Πειραματική αξιολόγηση του συστήματος.

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

Βιβλιογραφία

[1] K. Doka, D. Tsoumakos, and N. Koziris. Distributing the Power of OLAP. In *Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing (HPDC'10)*, pages 324–327. ACM, 2010. 2

[2] J. Gray, S. Chaudhuri, A. Bosworth, A. Layman, D. Reichart, M. Venkatrao, F. Pellow, and H. Pirahesh. Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals. *Data Min. Knowl. Discov.*, 1(1):29–53, 1997. 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