

Εργαστήριο Τεχνολογίας και Επεξεργασίας Δεδομένων

Θέματα Πτυχιακών Εργασιών 2011-2012

Επιβλέπων: Απόστολος Ν. Παπαδόπουλος

Στις επόμενες σελίδες περιγράφονται συνοπτικά τα θέματα πτυχιακών εργασιών για το ακαδημαϊκό έτος 2011-2012. Παρακαλώ όσοι ενδιαφέρονται να στείλουν μία αναλυτική βαθμολογία και τις προτιμήσεις για τα θέματα. Για την εκπόνηση των θεμάτων θεωρείται απαραίτητη η γνώση γλωσσών προγραμματισμού (C++ ή JAVA ανάλογα με το θέμα), και καλό υπόβαθρο σε μαθήματα όπως **δομές δεδομένων, θεωρία αλγορίθμων**. Επίσης, θα βοηθήσουν πάρα πολύ γνώσεις στα αντικείμενα των **βάσεων δεδομένων** και της **εξόρυξης δεδομένων**. Παρακαλούνται όσοι ενδιαφέρονται να επικοινωνήσουν με e-mail στο papadopo@csd.auth.gr μέχρι τις **30 Νοέμβρη 2011**. Τονίζεται ότι από τη στιγμή που γίνει η ανάθεση των θεμάτων αυτή **ισχύει μέχρι την εξεταστική του Σεπτεμβρίου 2012** οπότε και πρέπει να ολοκληρωθεί η εργασία.

Ανάλογα με το “βάθος” της μελέτης του θέματος, στις περισσότερες εργασίες υπάρχει η δυνατότητα να γίνει δημοσίευση σε πρακτικά συνεδρίων ή ακόμη και σε επιστημονικό περιοδικό. Απαραίτητη προϋπόθεση για αυτό είναι να υπάρχει όρεξη για δουλειά και το αποτέλεσμα της εργασίας να εμπεριέχει αρκετά στοιχεία πρωτοτυπίας.

Σημειώνεται επίσης ότι μετά την ολοκλήρωση της πτυσιακής θα πραγματοποιηθεί και παρουσίαση των αποτελεσμάτων στα μέλη του εργαστηρίου, οπότε θα απαιτηθεί και κάποιο αρχείο powerpoint.

Θέμα 1

Ερωτήματα κορυφογραμμής σε περιβάλλον MapReduce

Περιγραφή

Ένα ερώτημα κορυφογραμμής (skyline) λαμβάνει στην είσοδο ένα σύνολο πολυδιάστατων δεδομένων (π.χ. στις 2 διαστάσεις) και επιστρέφει το σύνολο των αντικειμένων (σημείων) που είναι τα καλύτερα δυνατά. Για παράδειγμα, έστω ένα σύνολο από φορητούς υπολογιστές για τους οποίους καταγράφουμε την τιμή και την ταχύτητα του επεξεργαστή. Οι καλύτεροι υπολογιστές είναι αυτοί που έχουν χαμηλή τιμή (μικρές τιμές της διάστασης “τιμή”) και μεγάλη ταχύτητα επεξεργαστή (μεγάλες τιμές της διάστασης “ταχύτητα”). Στην περίπτωση αυτή, οι καλύτεροι υπολογιστές προσδιορίζονται από μία διαδικασία βελτιστοποίησης με πολλαπλά κριτήρια.

Στόχος της πτυχιακής είναι η υλοποίηση μεθόδων προσδιορισμού της κορυφογραμμής ενός συνόλου δεδομένων σε καταναμημένο περιβάλλον και συγκεκριμένα σε MapReduce. Θα χρησιμοποιηθεί η open source έκδοση Hadoop για το σκοπό αυτό. Μετά την υλοποίηση θα πραγματοποιηθεί συγκριτική μελέτη των μεθόδων με τη βοήθεια μεγάλων συνόλων δεδομένων (πολλά Gbytes).

Λαμβάνοντας υπόψη ότι οι εργασίες που μελετούν ερωτήματα κορυφογραμμής σε MapReduce είναι ελάχιστες, η πτυχιακή μπορεί να οδηγήσει σε δημοσίευση αρκεί να υπάρχει όρεξη ώστε να μελετηθούν και νέοι αλγόριθμοι που λύνουν το πρόβλημα.

Απαιτήσεις

Καλή γνώση JAVA και η γνώση Hadoop θα είναι plus αλλά δεν είναι απαραίτητη (μπορεί να αποκτηθεί και κατά την εκπόνηση της εργασίας).

Χρήσιμο Υλικό

B. Zhang, S. Zhou, J. Guan, “Adapting Skyline Computation in the MapReduce Framework”, *DASFAA*, 2011. <http://delab.csd.auth.gr/~apostol/thesis1.pdf>

Θέμα 2

Αλγόριθμοι γραφημάτων σε περιβάλλον MapReduce

Περιγραφή

Πολλές από τις σύγχρονες εφαρμογές απαιτούν τη διαχείριση μεγάλων γραφημάτων (π.χ. Web, facebook, κλπ). Το βασικό πρόβλημα στις περιπτώσεις αυτές είναι ότι το μέγεθος των γραφημάτων είναι τεράστιο με αποτέλεσμα να δημιουργούνται προβλήματα στη διαχείρισή τους. Μία από τις εναλλακτικές είναι η χρήση πολλαπλών πόρων, όπως συμβαίνει στην περίπτωση του MapReduce.

Στόχος της πτυχιακής εργασίας είναι η μελέτη αλγορίθμων γραφημάτων σε περιβάλλον MapReduce και η εξαγωγή πειραματικών μετρήσεων χρησιμοποιώντας πολύ μεγάλα γραφήματα. Συγκεκριμένα, θα μελετηθεί **το πρόβλημα του υπολογισμού του αριθμού των τριγώνων σε ένα γράφημα τόσο συνολικά όσο και ανά κορυφή**. Επίσης, μας ενδιαφέρουν και προσεγγιστικοί αλγόριθμοι με κάποια εγγύηση όμως ως προς το λάθος.

Απαιτήσεις

Καλή γνώση JAVA και η γνώση Hadoop θα είναι plus αλλά δεν είναι απαραίτητη (μπορεί να αποκτηθεί και κατά την εκπόνηση της εργασίας).

Χρήσιμο Υλικό

S. Suri and S. Vassilvitskii, "Counting Triangles and the Curse of the Last Reducer", *WWW*, 2011.
<http://delab.csd.auth.gr/~apostol/thesis2.pdf>

Θέμα 3

Ανάπτυξη εφαρμογών σε κατανεμημένες βάσεις δεδομένων Cassandra ή Hbase (σε Hadoop)

Περιγραφή

Τον τελευταίο καιρό η κοινότητα των βάσεων δεδομένων αναζητεί τρόπους για γρηγορότερη επεξεργασία των δεδομένων έτσι ώστε να είναι δυνατή η διαχείριση πολλών Tbytes (ή ακόμη και Pbytes). Μία ενδιαφέρουσα κατεύθυνση με πολλές προοπτικές είναι η χρήση κατανεμημένων βάσεων δεδομένων. Αν και οι κατανεμημένες βάσεις δεδομένων έχουν πολύ μεγάλη ιστορία, στη σύγχρονη εποχή έχουν αλλάξει κατά πολύ τα δεδομένα τόσο από την πλευρά του hardware όσο και από την πλευρά των συστημάτων. Τα συστήματα Cassandra και Hbase είναι δύο κατανεμημένα NOSQL συστήματα που μπορούν να τρέξουν πάνω από το HDFS (Hadoop Distributed File System) προσφέροντας μεγάλες δυνατότητες κλιμάκωσης. Στην εργασία αυτή θα πρέπει αρχικά να εγκαταστήσετε την Hbase ή την Cassandra πάνω από Hadoop, να αναπτύξετε εφαρμογές και να δώσετε πειραματικά αποτελέσματα. Συγκεκριμένα, θα πρέπει να μπορούν να αποθηκευθούν πολλά Gbytes ή Tbytes δεδομένων (ανάλογα με τους διαθέσιμους servers) και στη συνέχεια να μπορούν να εκτελούνται ερωτήματα από κάποιον client.

Απαιτήσεις

Η γνώση Hadoop θα είναι ένα συν, αλλά δεν απαιτείται καθώς μπορεί να αποκτηθεί κατά την εκπόνηση της εργασίας. Το hadoop είναι ήδη στημένο σε servers του εργαστηρίου, αλλά αν θέλετε να δουλεύετε και ανεξάρτητα θα πρέπει να έχετε γνώσεις από ubuntu linux για την εγκατάσταση του Hadoop και της Hbase (ή της Cassandra).

Χρήσιμο Υλικό

<http://hadoop.apache.org>

Θέμα 4

Κατάταξη αντικειμένων χρησιμοποιώντας συσχετίσεις

Περιγραφή

Πολλές φορές ενδιαφερόμαστε για την αναζήτηση σχετικών αντικειμένων με βάση κάποιο ερώτημα που θέτει ο χρήστης. Ας αναφέρουμε ως παράδειγμα έναν πίνακα δεδομένων που αποθηκεύει τα χαρακτηριστικά κάποιων laptops ενώ σε έναν άλλο πίνακα αποθηκεύονται διάφορες κριτικές για τα συγκεκριμένα μοντέλα. Ένας χρήστης μπορεί να ενδιαφέρεται για την αγορά ενός laptop που να έχει μεγάλη οθόνη και να είναι ελαφρύ. Επομένως, το λογικό είναι να δώσει κάποιες λέξεις κλειδιά όπως “*lightweight*” και “*large screen*”. Το πρόβλημα είναι ότι το σύστημα δε θα μπορέσει να δώσει απάντηση διότι κανένα από αυτά τα keywords δεν εμφανίζεται στις περιγραφές των laptops. Όμως, οι λέξεις αυτές μπορεί να εμφανίζονται στις **κριτικές** για τα laptops. Άρα το πρόβλημα που καλούμαστε να λύσουμε εδώ είναι με ποιον τρόπο θα εκμεταλλευτούμε τις **συσχετίσεις μεταξύ των δεδομένων** ώστε να επιλέξουμε **τα καλύτερα δυνατά αντικείμενα** για το χρήστη.

Απαιτήσεις

Το σύστημα θα υλοποιηθεί σε γλώσσα JAVA ή C++.

Χρήσιμο Υλικό

K. Chakrabarti, V.Ganti, J. Han,

“Ranking Objects by Exploiting Relationships: computing top-k over aggregation”, *ACM SIGMOD*, 2006.

<http://delab.csd.auth.gr/~apostol/thesis4.pdf>

Θέμα 5

Έλεγχος ύπαρξης μονοπατιού σε πιθανοτικά γραφήματα

Περιγραφή

Πολλές σύγχρονες εφαρμογές χρειάζονται τη διαχείριση γραφημάτων (graphs). Στην εργασία αυτή θα έχετε την ευκαιρία να ασχοληθείτε με ένα αρκετά σημαντικό πρόβλημα στην περιοχή, που είναι το πρόβλημα της σύνδεσης μεταξύ δύο κόμβων (αν δηλαδή μπορούμε να πάμε από τον ένα κόμβο στον άλλο). Το γράφημα μπορεί να αναπαριστά ένα κοινωνικό δίκτυο, ένα δίκτυο υπολογιστών και γενικά οτιδήποτε μπορεί να αναπαρασταθεί με γράφημα. Δίνεται ένα κατευθυνόμενο γράφημα G . Σε κάθε ακμή του G αντιστοιχεί μία τιμή μεταξύ 0 και 1 που αναπαριστά την πιθανότητα ύπαρξης της συγκεκριμένης ακμής. Αν δίνεται ένα ζεύγος κορυφών x , y και ένα κατώφλι p , ο στόχος είναι να διαπιστώσουμε εάν υπάρχει μονοπάτι που να οδηγεί από την κορυφή x στην κορυφή y και να έχει πιθανότητα ύπαρξης μεγαλύτερη ή ίση από p .

Απαιτήσεις

Η υλοποίηση θα πραγματοποιηθεί σε C++ χρησιμοποιώντας την **lemon graph library** για τη δημιουργία και τη διαχείριση δεδομένων γραφημάτων. Τα πειραματικά αποτελέσματα θα βασιστούν σε πραγματικά σύνολα δεδομένων (π.χ. από SNAP repository).

Χρήσιμο Υλικό

<http://delab.csd.auth.gr/~apostol/thesis5.pdf>

<http://lemon.cs.elte.hu/trac/lemon>