ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ

ΤΜΗΜΑ ΕΠΙΣΤΗΜΗΣ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΑΡΟΥΣΙΑΣΗ / ΕΞΕΤΑΣΗ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

Αγαθάγγελος Ιωάννης
Μεταπτυχιακός Φοιτητής

Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης
Επόπτης Μεταπτ. Εργασίας: Καθηγητής, Δ. Πλεξουσάκης

Πέμπτη, 31/01/2019, 11:00
Αίθουσα Κ206,Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

" Κατακερματισμός δεδομένων RDF με σκοπό την αποτελεσματική απάντηση επερωτήσεων εκμεταλλευόμενοι την τοπικότητα των δεδομένων"

ΠΕΡΙΛΗΨΗ

Η διόγκωση του παγκόσμιου ιστού και η αφθονία των διασυνδεδεμένων δεδομένων, απαιτούν την ανάπτυξη αποτελεσματικών μεθόδων για την διαχείριση και την αποθήκευσή τους, καθώς και την αναζήτηση σε αυτά. Το Apache Spark είναι μία από τις πιο ενεργές και δημοφιλής πλατφόρμες μεγάλων δεδομένων, και πλέον όλο και περισσότερα συστήματα το χρησιμοποιούν για αποτελεσματική απάντηση επερωτήσεων σε κατανεμημένα δεδομένα. Μέχρι τώρα, οι περισσότερες δουλειές που βασίζονται στο Spark με σκοπό την απάντηση επερωτήσεων σε RDF δεδομένα, χρησιμοποιούν απλές τεχνικές οριζόντιου ή κάθετου κατακερματισμού. Αυτές οι τεχνικές συνήθως οδηγούν σε χαμηλή απόδοση όσον αφορά τον χρόνο εκτέλεσης των επερωτήσεων, καθώς δεν καταφέρνουν να αναγνωρίσουν την τοπικότητα των δεδομένων και συνεπώς να ομαδοποιήσουν σύνολα αυτών που συνήθως επερωτώνται μαζί.

Για να αντιμετωπίσουμε το πρόβλημα αυτό, σε αυτή την μεταπτυχιακή εργασία παρουσιάζουμε το LAWA , μία καινοτόμα πλατφόρμα που δέχεται σαν είσοδο ένα σύνολο δεδομένων RDF και είναι ικανή να τα κατακερματίσει αποδοτικά εξασφαλίζοντας την βέλτιστη τοπικότητα στα κατακερματισμένα μέρη. Για να το καταφέρουμε αυτό, αρχικά αναγνωρίζουμε τους πιο σημαντικούς κόμβους του συνόλου των δεδομένων σαν ”κεντρικούς κόμβους” και στην συνέχεια κατανείμουμε τους υπόλοιπους κόμβους στον ”κεντρικό κόμβο” που έχουν την μεγαλύτερη εξάρτηση. Η τεχνική αυτή εξασφαλίζει την τοπικότητα των δεδομένων και στις περισσότερες περιπτώσεις παράγει μία ομοιόμορφη κατανομή. Προκειμένου να μελετήσουμε περισσότερο τις σχεδιαστικές επιλογές απομονώνοντας τις επιπτώσεις της τοπικότητας και της κατανομής των δεδομένων, εισαγάγουμε δυο διαφορετικές τεχνικές. Η πρώτη εστιάζει στην τοπικότητα των δεδομένων (κατακερματισμός δεδομένων χρησιμοποιώντας την τοπικότητα αυτών) και η δεύτερη εγγυάται επιπλέον την ομοιόμορφη κατανομή των δεδομένων (οριοθετημένος κατακερματισμός δεδομένων χρησιμοποιώντας την τοπικότητα αυτών).Δείχνουμε ότι η προσέγγισή μας προσφέρει βέλτιστη ισορροπία μεταξύ της κατανομής των δεδομένων, του αριθμού αντιγράφων, και της μείωσης του όγκου δεδομένων που απαιτείται για την αποτίμηση επερωτήσεων, υπερτερώντας απέναντι σε υπάρχουσες προσεγγίσεις. Πιο συγκεκριμένα αξιολογήσαμε την δουλεία μας χρησιμοποιώντας συνθετικά και πραγματικά σύνολα δεδομένων, αποδεικνύοντας ότι βελτιώνουμε τον χρόνο εκτέλεσης των επερωτήσεων κατα τάξεις μεγέθους σε σύγκριση με τα μέχρι τώρα συστήματα της περιοχής.

Ioannis Agathaggelos
M.Sc. Thesis

Computer Science Department
University of Crete
Master's Thesis Supervisor: Professor, D. Pleksousakis

Thursday 31/01/2019, 11:00
Room K206, Computer Science Dept., University of Crete

“LAWA: Locality Aware Partitioning for Efficient Query Answering Over RDF Data”

ABSTRACT

The explosion of the web and the abundance of linked data, demand for effective and efficient methods for storage, management and querying. Apache Spark is one of the most active big-data approaches, with more and more systems adopting it, for efficient querying over distributed data. However, most of the Spark-based RDF query answering approaches so far, are exploiting simplistic horizontal and/or vertical partitioning of triples, resulting in poor query performance. The main reason for this is that simplistic data partitioning approaches fail to identify data locality and group together data that are usually queried together.

To mitigate this problem, in this thesis, we present LAWA, a novel platform that accepts as input an RDF dataset and effectively partitions data, ensuring data locality. To achieve this, we identify the dataset's important nodes as centroids and then we distribute the other nodes to the centroid they mostly depend on.

This scheme ensures data locality, and in most cases results in a balanced data distribution, however, without offering any guarantees on it. In order to study the design choices and isolate the impact of data locality and data distribution we introduce two variants. One focusing purely only on data locality, i.e. the Locality Aware Partitioning (LAP) approach and another one enforcing balanced data distribution as well, i.e. the Bounded Locality Aware Partitioning (BLAP) approach.

We show that out approach offers an optimal fine tuning between data distribution, replication and data reduction, dominating existing approaches. More specifically we evaluate our approach using both synthetic and real workloads, showing that we improve query answering orders of magnitude over existing state of the art.