FlexRecs: Εκφράζοντας και Συνδυάζοντας Ευέλικτες Συστάσεις

Γεωργία Κούτρικα, Benjamin Bercovitz, Hector Garcia-Molina

Τα συστήματα συστάσεων έχουν γίνει πολύ δημοφιλή, παρόλα αυτά οι περισσότερες μέθοδοι συστάσεων είναι ενσωματωμένες στο σύστημα γεγονός που δυσχεραίνει τον πειραματισμό και την υλοποίηση νέων μεθόδων συστάσεων. Στο παρόν άρθρο, προτείνουμε ένα πλαίσιο που αποσυνδέει την περιγραφή μίας μεθόδου από την εκτέλεσή της και υποστηρίζει ευέλικτες συστάσεις πάνω από δομημένα δεδομένα. Μία μέθοδος συστάσεων περιγράφεται δηλωτικά ως μία παραμετροποιήσιμη ροή υψηλού επιπέδου που αποτελείται από κλασσικούς σχεσιακούς τελεστές και νέους τελεστές για τη δημιουργία και σύνθεση συστάσεων. Περιγράφουμε μία μηχανή ευέλικτων συστάσεων που υλοποιεί το παρόν πλαίσιο και περιγράφουμε παραδείγματα ροών καθώς και πειραματικά δεδομένα που δείχνουν τις δυνατότητες για σύλληψη και υλοποίηση πολλαπλών, παλιών ή καινοτόμων, μεθόδων συστάσεων εύκολα και αποδοτικά.

Συστήματα Θέσπισης για Εκτέλεση Συνεχών Ροών Εργασίας

Παναγιώτης Νεοφύτου, Πάνος Χρυσάνθης, Αλέξανδρος Λαμπρινίδης

Παραδοσιακά τα Συστήματα Θέσπισης "Ροών Εργασίας" (PE) (workflows) καθώς επίσης και συστήματα σχεδιασμού PE θεωρούν τις PE ως αλληλεπιδράσεις μιας φοράς με τις διάφορες πηγές δεδομένων. Κατά συνέπεια οι PE εκτελούν μια σειρά από βήματα μια φορά, οποτεδήποτε υπάρχει ανάγκη για επεξεργασία δεδομένων και παραγωγή αποτελεσμάτων. Η θεμελιώδης υποκειμενική υπόθεση μέχρι τώρα ήταν ότι οι πηγές δεδομένων είναι παθητικές και όλες οι αλληλεπιδράσεις με αυτές είναι δομημένες γύρω από τις γραμμές του μοντέλου ερωτημάτων (query). Άρα, τα παραδοσιακά Συστήματα Θέσπισης Ροών Εργασίας δεν έχουν την δυνατότητα υποστήριξης Επιστημονικών ή Επιχειρησιακών εφαρμογών παρακολούθησης, οι οποίες απαιτούν την κατεργασία ροών δεδομένων (data streams). Στα πλαίσια της δουλειάς αυτής, προτείνουμε μια αναθεώρηση του μοντέλου PE από το παραδοσιακό σταδιακό/βήμα-βήμα μοντέλο εκτέλεσης ροών εργασίας προς ένα μοντέλο συνεχούς εκτέλεσης των βημάτων, ούτως ώστε να καταστεί δυνατός ο χειρισμός ροών δεδομένων, προερχομένων και παραδοτέων ασύγχρονα από διαφορετικές πηγές.

Σχεδίαση ETL Διεργασιών για Βελτιστοποίηση Μετρικών Ποιότητας QoX

Άλκης Σιμιτσής, Kevin Wilkinson, Malu Castellanos, Umeshwar Dayal

Η επιχειρησιακή νοημοσύνη (Business Intelligence - BI) αποτελεί σημαντικό εργαλείο για το στρατηγικό σχεδιασμό σε οργανισμούς και επιχειρήσεις και στηρίζεται σε μεγάλο βαθμό σε διεργασίες Εξαγωγής-Μετασχηματισμού-Φόρτωσης δεδομένων (Extraction-Transformation-Loading - ETL processes) που ενημερώνουν κέντρα συγκέντρωσης πληροφοριών, όπως οι αποθήκες δεδομένων, συγκεντρώνοντας πληροφορίες από διάφορες διασκορπισμένες και ετερογενείς πηγές δεδομένων. Εξαιτίας των αυξημένων επιχειρησιακών αναγκών και των μειωμένων χρονικών ορίων που δίνονται για την εκτέλεση των διεργασιών ΕΜΦ, η σχεδίαση αυτών είναι εξαιρετικά πολύπλοκη, ενώ η βελτίωση της αποδοτικότητάς τους είναι ιδιαίτερα κρίσιμη.

Καθώς η επιχειρησιακή νοημοσύνη προσαρμόζεται σε σύγχρονες ανάγκες, αποκτά και λειτουργικό χαρακτήρα πέρα από τον παραδοσιακό στρατηγικό ρόλο που συνεχίζει να διαδραματίζει. Η εξέλιξη αυτή περιπλέκει τη βελτιστοποίηση της σχεδίασης διεργασιών ΕΜΦ, που πλέον πρέπει να ικανοποιεί ένα πλήθος μετρικών ποιότητας σε αρμονία με την αποδοτική εκτέλεση των διεργασιών αυτών. Παραδείγματα τέτοιων μετρικών αποτελούν η αξιοπιστία, η ευελιξία, η ελεγξιμότητα, καθώς και η ικανότητα των διεργασιών ΕΜΦ να ανταποκρίνονται σε τυχαίες μεταβολές της σχεδίασης, σε σφάλματα εκτέλεσης, σε απαιτήσεις για ελαχιστοποίηση καθυστερήσεων ενημέρωσης, κ.ο.κ.. Αναφερόμαστε

συγκεντρωτικά σε τέτοιες μετρικές ποιότητας με τον όρο μετρικές QoX. Οι υπάρχουσες τεχνικές σχεδίασης και βελτιστοποίησης δε λαμβάνουν υπόψη τέτοια ποιοτικά κριτήρια.
Σε αυτήν την εργασία πραγματευόμαστε θέματα βελτιστοποίησης διεργασιών ΕΜΦ με βάση μετρικές QoX σε όλα τα στάδια σχεδίασης: εννοιολογικό, λογικό και φυσικό. Περιγράφουμε ευριστικές τεχνικές σχεδίασης για ένα υποσύνολο των μετρικών QoX, όπως η απόδοση, η ανάκαμψη, η καθυστέρηση ενημέρωσης, και η αξιοπιστία διεργασιών ΕΜΦ. Δυστυχώς, η ικανοποίηση των απαιτήσεων ως προς μία μετρική ενδέχεται να δυσκολέψει την ικανοποίηση μιας άλλης μετρικής. Αναδεικνύουμε το πρόβλημα και πραγματευόμαστε τεχνικές ικανοποίησης απαιτήσεων που αφορούν σε περισσότερες από μία μετρικές QoX, χρησιμοποιώντας ένα τμήμα πραγματικού σεναρίου ΕΜΦ.

Βελτιστότητα και Κλιμακωσιμότητα στην Κατασκευή Ιστογραμμάτων Πλέγματος

Παναγιώτης Καρράς

Το Ιστόγραμμα Πλέγματος είναι μια προσφάτως προταθείσα τεχνική συνόψεως δεδομένων που επιτυγχάνει προσεγγιστική ποιότητα προτιμότερη από εκείνη ενός βελτίστου απλού ιστογράμματος. Όπως και άλλες μέθοδοι ιεραρχικής συνόψεως, ένα ιστόγραμμα πλέγαμτος (ΙΠ) αποσκοπεί να προσεγγίσει δεδομένα χρησιμοποιώντας μια ιεραρχική δομή. Ωστόσο, η δομή αυτή δεν ορίζεται εκ των προτέρων - συνιστά ένα ζητούμενο, όχι ένα δεδομένο, του προβλήματος. Προγενέστερη έρευνα έχει καθορίσει τις ιδιότητες που πρέπει ένα ΙΠ να υπακούει και ανέπτυξε προσεγγιστικούς αλγορίθμους γενικής χρήσεως για την κατασκευή του. Ωστόσο, δύο σημαντικά ζητήματα παραμένουν ανεξέταστα: Πρώτον, η κατασκευή ενός βελτίστου ΙΠ για ένα καθορισμένο μέτρο σφάλματος είναι ένα πρόβλημα άλυτο μέχρι σήμερα. Δεύτερον, οι προταθέντες αλγόριθμοι πάσχουν από πολύ υψηλές πολυπλοκότητες χώρου και χρόνου που καθιστούν την εφαρμογή τους σε συνθήκες πραγματικού κόσμου προβληματική. Σε αυτό το άρθρο, εξετάζουμε και τα δύο αυτά ζητούμενα, με επίκεντρο την περίπτωση που το στοχούμενο μέτρο σφάλματος είναι ένα μέτρο μεγίστου σφάλματος. Οι αλγόριθμοί μας αντιμετωπίζουν τόσο το πρόβλημα κατασκευής ΙΠ με οριοθετούμενο σφάλμα, στο οποίο ο χώρος που καταλαμβάνεται από ένα ΙΠ ελαχιστοποιείται υπό ένα όριο σφάλματος, καθώς και το κλασικό πρόβλημα οριοθετούμενου χώρου. Πρώτα αναπτύσσουμε έναν αλγόριθμο δυναμικού προγραμματισμού που ανιχνεύει ένα βέλτιστο ΙΠ υπό καθορισμένο όριο μεγίστου σφάλματος. Έπειτα προτείνουμε έναν αποτελεσματικό, πρακτικό, άπληστο αλγόριθμο που λύνει το ίδιο πρόβλημα με πολύ χαμηλότερες απαιτήσεις χώρου και χρόνου. Στη συνέχεια, δείχνουμε ότι και οι δύο μας αλγόριθμοι μπορούν να εφαρμοστούν στο κλασικό πρόβλημα οριοθετούμενου χώρου, που στοχεύει στην ελαχιστοποίηση σφάλματος υπό ένα όριο χώρου. Η πειραματική μας μελέτη με δεδομένα πραγματικού κόσμου δείχνει την αποτελεσματικότητα των μεθόδων μας σε σύγκριση με ανταγωνιστικές τεχνικές συνόψεως. Επιπλέον, τα πορίσματά μας δείχνουν ότι η άπληστη ευρετική μας αποδίδει σχεδόν το ίδιο καλά με τη βέλτιστη λύση σε ακρίβεια.

Τυχαίες εγγραφές σε μνήμες : ανακτώντας το χαμένο χρόνο.

Radu Stoica, Μάνος Αθανασούλης, Ryan Johnson, Αναστασία Αϊλαμάκη

Η τεχνολογία συμπαγούς κατάστασης flash προσφέρει γρηγορότερη προσπέλαση στα δεδομένα σε σύγκριση με τους μαγνητικούς σκληρούς δίσκους, αλλά η τιμή της ήταν ως τώρα απαγορευτικά υψηλή για χρήση σε συστήματα διαχείρισης βάσεων δεδομένων (ΣΔΒΔ). Τα τελευταία χρόνια, όμως, οι μνήμες flash γίνονται όλο και πιο δημοφιλείς γιατί η τιμή ανά GB μειώνεται δραστικά και η διαθέσιμη χωρητικότητα αυξάνεται εκθετικά με το χρόνο. Τα ΣΔΒΔ έχουν αρχίσει να χρησιμοποιούν μνήμες flash είτε αυτόνομα είτε σε υβριδικά σχήματα με χρήση μαγνητικών δίσκων. Οι μνήμες flash, όμως, συμπεριφέρονται διαφορετικά από τους μαγνητικούς δίσκους σε κοινές αλληλουχίες προσπελάσεων. Πρόσφατα ερευνητικά αποτελέσματα αναδεικνύουν την ασυμμετρία μεταξύ των ταχυτήτων εγγραφής

και ανάγνωσης, καθώς και τη μη προβλεψιμότητα της απόδοσης της μνήμης flash, η οποία επηρεάζεται ισχυρά από το ιστορικό των προσπελάσεων.

Σε αυτήν την εργασία αξιολογούμε τη σημασία των τυχαίων εγγραφών σε μνήμες flash, οι οποίες αποδεικνύονται πολύ πιο ακριβές από τις σειριακές εγγραφές, ένα χαρακτηριστικό που δεν επιτρέπει στα ΣΔΒΔ να εκμεταλλευτούν τις μνήμες flash, οι οποίες λόγω της έλλειψης μηχανικών μερών, έχουν πολύ μικρό κόστος για τις τυχαίες αναγνώσεις. Για να εκμεταλλευτούμε πλήρως της μνήμες flash μετατρέπουμε τις τυχαίες εγγραφές σε σειριακές, κάτι που έχει ως αποτέλεσμα περισσότερες σειριακές εγγραφές και τυχαίες αναγνώσεις, οι οποίες, όμως, λόγω του πολύ μικρότερου - από τις τυχαίες εγγραφές - κόστους τους, οδηγούν σε συνολική βελτίωση της απόδοσης της μνήμης flash. Προτείνουμε έναν αλγόριθμο ο οποίος προσαρτεί σειριακά τις σελίδες όταν ανανεώνονται τα περιεχόμενά τους και ανακτά περιοδικά το χώρο στον οποίο εδράζουν σελίδες με παρωχημένα δεδομένα. Αξιολογούμε τον προτεινόμενο αλγόριθμο με χρήση φόρτου εργασίας τύπου συστημάτων διεκπεραίωσης συναλλαγών (OLTP), χρησιμοποιώντας τόσο μία αναλυτική πρόβλεψη μοντελοποιώντας τη συμπεριφορά του αλγορίθμου και της μνήμης, όσο και πειράματα αναφοράς. Τα αποτελέσματα δείχνουν ότι λαμβάνοντας υπ'όψιν τις ιδιαιτερότητες της μνήμης flash, μπορούμε να πετύχουμε σημαντική επιτάχυνση στην απόδοση εγγραφής σε ρεαλιστικό φόρτο εργασίας.

Βελτιστοποίηση ρυθμού παραγωγής αποτελέσματος σύζευξης σε δυναμικά περιβάλλοντα: Διπλά ευρετηριασμένη σύζευξη φωλιασμένου βρόχου με ανάδραση

Μιχαέλα Βορνέα, Βασίλης Βασάλος, Γιάννης Κωτίδης, Αντώνιος Δεληγιαννάκης

Οι προσαρμοστικοί αλγόριθμοι σύζευξης έχουν συγκεντρώσει πολύ ερευνητικό ενδιαφέρον τα τελευταία χρόνια, ειδικά σε εφαρμογές όπου τα δεδομένα παρέχονται από αυτόνομες πηγές μέσω ετερογενών και όχι πάντα προβλέψιμων δικτυακών υποδομών. Το βασικό τους πλεονέκτημα σε σχέση με παραδοσιακούς αλγορίθμους σύζευξης είναι ότι μπορούν να αρχίσουν να παράγουν πλειάδες σχεδόν με την παραλαβή των πρώτων πλειάδων εισόδου, με αποτέλεσμα τη βελτίωση της σωλήνωσης του αλγορίθμου. Η βελτίωση αυτή προκαλείται από την εξομάλυνση του ρυθμού παραγωγής αποτελεσμάτων και την κάλυψη τυχόν καθυστερήσεων στο δίκτυο ή τις πηγές δεδομένων. Στο άρθρο αυτό προτείνουμε τον αλγόριθμο Διπλά Ευρετηριασμένης Σύζευξης Φωλιασμένου Βρόχου με Ανάδραση, ένα νέο προσαρμοστικό αλγόριθμο για μεγιστοποίηση ρυθμού παραγωγής αποτελέσματος. Ο αλγόριθμός μας (τα αρχικά στα αγγλικά δίνουν το αρτικόλεξο DINER) συνδυάζει δυο βασικά στοιχεία: α) Μια διαισθητικά (και πειραματικά) αποτελεσματική πολιτική αντικατάστασης ενταμιευτών που έχει στόχο την αύξηση της παραγωγικότητας των πλειάδων που αποθηκεύονται στους ενταμιευτές (δηλαδή την αύξηση της συμμετοχής τους στην παραγωγή αποτελεσμάτων σύζευξης) και β) Μια νέα τεχνική σύζευξης με επανείσοδο (reentrant join technique) που επιτρέπει στον αλγόριθμο να αλλάζει γρήγορα μεταξύ επεξεργασίας πλειάδων στη μνήμη και αυτών που έχουν σωθεί στο δίσκο, έτσι ώστε να εκμεταλλεύεται αποτελεσματικότερα τις καθυστερήσεις αφίξεων πλειάδων. Η πειραματική μας μελέτη με πραγματικά και συνθετικά δεδομένα δείχνει ότι ο DINER έχει πολύ καλύτερες επιδόσεις από τους υπάρχοντες προσαρμοστικούς αλγορίθμους σε ότι αφορά το ρυθμό παραγωγής πλειάδων της σύζευξης, ενώ αξιοποιεί καλύτερα τη διαθέσιμη μνήμη.

Ένας αποδοτικός και προβλέψιμος τελεστής ζεύξης για αποθήκες δεδομένων με πολλά ταυτόχρονα ερωτήματα

Νεοκλής Πολυζώτης, George Candea, Radek Vingralek

Οι συμβατικές αποθήκες δεδομένων χρησιμοποιούν το μοντέλο "ένα ερώτημα την φορά", συσχετίζοντας κάθε ερώτημα με ένα ξεχωριστό πλάνο εκτέλεσης. Αυτή η προσέγγιση προξενεί συμφόρηση όταν πολλά ερωτήματα τρέχουν ταυτόχρονα, καθώς τα πλάνα συναγωνίζονται για την χρήση των πόρων εισόδου/εξόδου και υπολογισμού. Έτσι, ενώ τα μοντέρνα συστήματα μπορούν να απαντήσουν αποδοτικά ένα ερώτημα, η απόδοσή τους πέφτει αισθητά όταν υπάρχουν πολλά περίπλοκα ερωτήματα που αποτιμούνται ταυτόχρονα.

Αυτή η εργασία εισάγει μια επέκταση των συμβατικών συστημάτων βάσεων δεδομένων, που βελτιώνει το συνολικό ρυθμό έργου για ζεύξεις σε αποθήκες δεδομένων με πολλά ταυτόχρονα ερωτήματα. Σε αντίθεση με τον μοντέλο "ένα ερώτημα την φορά", η προσέγγισή μας χρησιμοποιεί ένα πλάνο εκτέλεσης που υλοποιεί την είσοδο/έξοδο, τον υπολογισμό ζεύξεων και την αποθήκευση πλειάδων για όλα τα ταυτόχρονα ερωτήματα μαζί, ώστε η συνολική δουλειά να "μοιράζεται" ουσιαστικά από όλα τα ερωτήματα. Το πλάνο είναι συνεχώς ενεργό, απαρτίζεται από μια αλληλουχία διασυνδεστικών τελεστών, και ελέγχεται από μία μονάδα λογισμικού που εξετάζει συνέχεια την λειτουργία του πλάνου και εφαρμόζει κατάλληλες βελτιστοποιήσεις. Η προτεινόμενη σχεδίαση επιτρέπει την αποδοτική αποτίμηση ερωτημάτων σε μεγάλο όγκο δεδομένων, με προβλέψιμους χρόνους εκτέλεσης και χαμηλή συμφόρηση. Τα πειραματικά αποτελέσματα δείχνουν ότι η προσέγγισή μας ξεπερνά κατά μία τάξη μεγέθους την απόδοση συμβατικών εμπορικών συστημάτων, όταν αποτιμούνται ταυτόχρονα δεκάδες με εκατοντάδες ερωτήματα.

Εκμετάλλευση της δύναμης των σχεσιακών βάσεων δεδομένων για την αποδοτική επεξεργασία δυναμικών ρευμάτων δεδομένων

Εριέττα Λιάρου, Martin Kersten, Romulo Goncalves, Στράτος Ιδρέος

Οι εφαρμογές δυναμικών ρευμάτων δεδομένων κέρδισαν σημαντική δημοτικότητα τα τελευταία χρόνια, γεγονός που συνέβαλλε στην ανάπτυξη εξειδικευμένων μηχανών για την επεξεργασία τους. Αυτά τα συστήματα σχεδιάζονται από την αρχή με μια διαφορετική φιλοσοφία συγκριτικά με τα παραδοσιακά συστήματα διαχείρισης βάσεων δεδομένων προκειμένου να αντιμετωπιστούν τις εξειδικευμένες απαιτήσεις των εφαρμογών αυτών. Εντούτοις, αυτό σημαίνει ότι στερούνται τη δύναμη, τις εξελιγμένες τεχνικές και τους αλγορίθμους που έχουν συσσωρευτεί κατά τη διάρκεια πολλών ετών έρευνας στις βάσεις δεδομένων.

Σε αυτή τη δουλειά, παίρνουμε την αντίθετη κατεύθυνση και σχεδιάζουμε μια μηχανή διαχείρισης δυναμικών ρευμάτων δεδομένων πάνω από έναν πυρήνα διαχείρισης δεδομένων. Οι εισερχόμενες πλειάδες αποθηκεύονται άμεσα κατά την άφιξη τους σε ένα νέο είδος πινάκων, τα αποκαλούμενα "καλάθια". Ένα ερώτημα διαρκείας που έχει υποβληθεί στο σύστημα, μπορεί να απαντηθεί όταν τα σχετικά προς αυτό καλάθια εμπεριέχουν νέες εισερχόμενες πλειάδες, σαν να ήταν ένα τυπικό ερώτημα που υποβάλλεται σε μια βάση δεδομένων. Μόλις μια πλειάδα γνωστοποιηθεί σε όλες τις σχετικές ερωτήσεις διαρκείας, διαγράφεται από το καλάθι της. Ένα καλάθι μπορεί να αποτελεί είσοδο ενός ή πολλαπλών ερωτήσεων. Επιπλέον, ένα προσχέδιο εκτέλεσης ενός ερωτήματος διαρκείας μπορεί να χωριστεί σε πολλαπλά μέρη, καθένα με τα δικά του καλάθια εισόδου-εξόδου, ώστε να επιτρέπουν καλύτερο προγραμματισμό εκτέλεσης. Σε αντίθεση με τις παραδοσιακές μηχανές διαχείρισης ρευμάτων δεδομένων, που επεξεργάζονται μια πλειάδα τη φορά, το προτεινόμενο πρότυπο επιτρέπει την επεξεργασία κατά δέσμες πλειάδων, π.χ., εκτέλεσε ένα ερώτημα μόνο αφού στα καλάθια εισόδου του έχουν έρθει Χ πλειάδες.

Σε αυτό το άρθρο, ερευνούμε τις ευκαιρίες και τις προκλήσεις που προκύπτουν με μια τέτοια κατεύθυνση και δείχνουμε ότι μπορεί να επιφέρει σημαντικά πλεονεκτήματα. Προτείνουμε μια πλήρη αρχιτεκτονική, το DataCell, το οποίο υλοποιούμε πάνω από ένα column-store. Επίσης παραθέτουμε λεπτομερή ανάλυση και πειραματική αξιολόγηση των αλγορίθμων που προτείνουμε.

Τυπικό Μοντέλο Αναπαράστασης Γνώσης Περιοχής και Εξαγωγής Συμπερασμάτων βάσει MPEG-7 Δομών

Χρύσα Τσιναράκη, Σταύρος Χριστοδουλάκης

Η χρήση γνώσης περιοχής στις περιγραφές βάσει σημαντικής του οπτικοακουστικού υλικού ενισχύει τη λειτουργικότητα και την αποτελεσματικότητα των εφαρμογών πολυμέσων. Το MPEG-7 όμως, που είναι το κυρίαρχο πρότυπο για την περιγραφή οπτικοακουστικού υλικού, δεν παρέχει μηχανισμούς για τη συστηματική ενσωμάτωση γνώσης περιοχής στις MPEG-7 περιγραφές και για την υποστήριξη εξαγωγής συμπερασμάτων βάσει των MPEG-7 περιγραφών. Ο ορισμός ενός τυπικού μοντέλου αναπαράστασης γνώσης περιοχής με τη χρήση MPEG-7 δομών, που θα επιτρέπει την εξαγωγή συμπερασμάτων βάσει των MPEG-7 περιγραφών, είναι ύψιστης σημασίας για την αξιοποίηση της γνώσης περιοχής ώστε να μπορεί να υποστηριχθεί η επεξεργασία του οπτικοακουστικού υλικού βάσει σημαντικής. Σε αυτό το άρθρο παρουσιάζουμε ένα τυπικό μοντέλο που επιτρέπει τη συστηματική ενσωμάτωση γνώσης περιοχής με τη χρήση MPEG-7 δομών και την αξιοποίησή της κατά την εξαγωγή συμπερασμάτων. Το προτεινόμενο μοντέλο αξιοποιεί αποκλειστικά MPEG-7 δομές, ενώ οι περιγραφές που δομούνται βάσει του μοντέλου είναι απολύτως συμβατές με το MPEG-7

Εξερευνητική Αναζήτηση του Ιστού μέσω Δυναμικών Ταξινομιών και Ομαδοποίησης Αποτελεσμάτων

Παναγιώτης Παπαδάκος, Στέλλα Κοπιδάκη, Νίκος Αρμενατζόγλου, Γιάννης Τζίτζικας

Η παρούσα εργασία προτείνει την αξιοποίηση των μεταδεδομένων (είτε δυναμικών ή στατικών) για τον εμπλουτισμό της διαδικασίας αναζήτησης στον Ιστό με υπηρεσίες εξερεύνησης. Η ομαδοποίηση αποτελεσμάτων σε πραγματικό χρόνο είναι χρήσιμη αφενός γιατί προσφέρει στους χρήστες μια εποπτική εικόνα των αποτελεσμάτων αναζήτησης και αφετέρου γιατί τους επιτρέπει να εστιάσουν στα επιθυμητά υποσύνολα των αποτελεσμάτων. Ονομάζουμε αυτά τα μεταδεδομένα δυναμικά αφού εξαρτώνται από την τρέχουσα επερώτηση. Από την άλλη μεριά, τα διάφορα μεταδεδομένα που είναι εκ των προτέρων διαθέσιμα στις μηχανές αναζήτησης, και άρα στατικά (π.χ. φυσική γλώσσα/ημερομηνία/τύπος στοιχείου απάντησης), συνήθως αξιοποιούνται από υπηρεσίες προηγμένης αναζήτησης (βασισμένες σε φόρμες) οι οποίες είναι δύσχρηστες και οι χρήστες σπάνια τις χρησιμοποιούν. Το άρθρο προτείνει μια προσέγγιση που συνδυάζει αμφότερους τύπους μεταδεδομένων μέσω του αλληλεπιδραστικού υποδείγματος των πολυδιάστατων δυναμικών ταξινομιών. Ο συνδυασμός αυτός προσφέρει μια αποτελεσματική, ευέλικτη και αποδοτική εξερευνητική εμπειρία.

Γενικής χρήσεως ανανεώσιμοι δείκτες για την εύρεση τιμών σε XML κείμενα

Λευτέρης Σιδηρουργός, Peter Boncz

Παρουσιάζουμε μια συλλογή από δείκτες για οποιοδήποτε τύπου κόμβου ενός XML κειμένου. Οι δείκτες αυτοί καταναλώνουν ελάχιστο χώρο, ανανεώνονται γρήγορα, υποστηρίζουν γρήγορη ανεύρεση αλφαριθμητικών τιμών καθώς και αριθμητικών διαστημάτων. Το δομικά στοιχεία των δεικτών είναι συναρτήσεις κατακερματισμού και μηχανές πεπερασμένων καταστάσεων. Τέλος, παρουσιάζουμε πειραματικά αποτελέσματα και συμπεράσματα από την υλοποίηση αυτών των δεικτών στην βάση δεδομένων MonetDB/XQuery.

Περιμετρική Ανάκτηση και Αντιγραφή Δεδομένων σε Κινητά Δίκτυα Αισθητήρων

Παναγιώτης Ανδρέου, Δημήτριος Ζεϊναλιπούρ-Γιαζτή, Μαρία Ανδρέου, Πάνος Χρυσάνθης, Γιώργος Σαμάρας

Αυτό το άρθρο υποθέτει ένα σύνολο από Ν κινητούς αισθητήρες οι οποίοι κινούνται σαν σμήνος σε ένα Ευκλείδειο επίπεδο. Οι στόχοι μας είναι να εξερευνήσουμε μία γεωγραφική περιοχή με την ανίχνευση χωρό-χρονικών γεγονότων που μας ενδιαφέρουν και ακολούθως να αποθηκεύσουμε αυτά τα δεδομένα τοπικά στους αισθητήρες του δικτύου μέχρι να τα ζητήσει ο χρήστης. Ένα τέτοιο πλαίσιο εργασίας βρίσκει εφαρμογή σε κινητά δίκτυα αισθητήρων στα οποία δεν υπάρχει συχνή επαφή του δικτύου με τον κεντρικό κόμβο (sink). Το πλαίσιο εργασίας μας, SenseSwarm, κατατάσσει τους αισθητήρες σε περιμετρικούς και εσωτερικούς αισθητήρες όπου οι πρώτοι είναι υπεύθυνοι για την ανίχνευση καινούριων γεγονότων ενώ οι δεύτεροι για την ασφαλή αποθήκευση των δεδομένων για μελλοντική χρήση.

Για την αποτελεσματική ανίχνευση των περιμετρικών αισθητήρων, έχουμε προτείνει τον Αλγόριθμο Περιμέτρου (Perimeter Algorithm - PA), ο οποίος είναι ένας αποδοτικός κατανεμημένος αλγόριθμος με χαμηλή πολυπλοκότητα μηνυμάτων. Για την ασφαλή αντιγραφή των δεδομένων έχουμε προτείνει τον Αλγόριθμο Αντιγραφής Δεδομένων (Data Replication Algorithm - DRA), ο οποίος είναι βασισμένος σε πλάνα αντιγραφής με ψήφους και επιτυγχάνει την ακριβή ανάκτηση των γεγονότων από το δίκτυο σε περιπτώσεις σφαλμάτων. Η πειραματική μας αξιολόγηση δείχνει ότι το πλαίσιο SenseSwarm παρέχει σημαντική μείωση κατανάλωσης ενέργειας προσφέροντας παράλληλα ένα ψηλό ρυθμό διαθεσιμότητας δεδομένων. Συγκεκριμένα, οι πειραματικές μας μετρήσεις δείχνουν ότι όταν τα σφάλματα σε ένα δίκτυο είναι κάτω από 60%, τότε μπορούμε να ανακτήσουμε επιτυχώς πέραν του 80% των μετρήσεων του δικτύου.

Συλλογισμός και αποτίμηση επερωτήσεων στο μοντέλο RDFS πάνω από κατανεμημένους πίνακες κατακερματισμού

Ζωή Καούδη, Ίρις Μηλιαράκη, Μανώλης Κουμπαράκης

Μελετάμε το πρόβλημα του κατανεμημένου συλλογισμού και της αποτίμησης επερωτήσεων στο μοντέλο RDFS πάνω από κατανεμημένους πίνακες κατακερματισμού. Ο κλιμακωτός και κατανεμημένος συλλογισμός RDFS είναι μια απαραίτητη λειτουργικότητα για τη βελτίωση της κλιμάκωσης και των επιδόσεων εφαρμογών του Σημασιολογικού Ιστού. Ο στόχος μας σε αυτήν την εργασία είναι να συγκρίνουμε και να αποτιμήσουμε δύο πολύ γνωστούς αλγορίθμους στη συλλογιστική, τους προς τα πίσω αλυσίδα εκτέλεσης (backward chaining) και προς τα εμπρός αλυσίδα εκτέλεσης (forward chaining), πάνω από κατανεμημένους πίνακες κατακερματισμού. Δείχνουμε πώς υλοποιούνται αυτοί οι δύο αλγόριθμοι πάνω από τον κατανεμημένο πίνακα κατακερματισμού Bamboo και αποδεικνύουμε την ορθότητά τους. Επίσης, μελετάμε τα πλεονεκτήματα/μειονεκτήματα των αλγορίθμων σχετικά με το χρόνο και το χώρο αναλυτικά μέσω ενός αναλυτικού μοντέλου, καθώς και πειραματικά αποτιμώντας τους αλγορίθμους στο PlanetLab.

Ταχεία και Αποδοτική Μείωση Διάστασης με Χρήση Σημείων Αναφοράς

Παναγής Μαγδαληνός, Χρήστος Δουλκερίδης, Μιχάλης Βαζιργιάννης

Η πρόοδος της επιστήμης σε διάφορα γνωσιακά πεδία έχει οδηγήσει σε πληροφοριακή υπερφόρτωση. Οι επιστήμονες έρχονται αντιμέτωποι με ολοένα και αυξανόμενου μεγέθους πειραματικά αποτελέσματα τα οποία καλούνται να ερμηνεύσουν και να αξιολογήσουν. Τα συγκεκριμένα σύνολα, παρουσιάζουν μια σειρά από προκλήσεις καθώς κλασικές μέθοδοι στατιστικής ανάλυσης αποτυγχάνουν να τα χειριστούν μιας και οι μεταβλητές που τα ορίζουν (οι διαστάσεις του πειράματος) έχουν αυξηθεί δραματικά. Ένα από τα βασικά προβλήματα που παρουσιάζουν τα πολυδιάστατα δεδομένα είναι η επιδείνωση της ποιότητας των ερωτήσεων, ένα φαινόμενο που χαρακτηρίζεται ως "κατάρα των πολλών διαστάσεων".

Στόχος της μεθοδολογίας μείωσης δεδομένων είναι η μείωση των διαστάσεων των δεδομένων με ταυτόχρονη διατήρηση των ιδιοτήτων εκείνων που θα επιτρέψουν την διενέργεια ποιοτικής ανάλυσης. Μεταφράζοντας την τελευταία απαίτησή μας σε μαθηματικό μοντέλο, δεδομένου ενός διανύσματος διάστασης n, αναζητούμε μια προβολή στο χώρο διάστασης k, με k<<n, η οποία διατηρεί το περιεχόμενο των αρχικών δεδομένων με βάση κάποιο συγκεκριμένο κριτήριο. Παράλληλα ο όγκος και η διάσταση των σύγχρονων δεδομένων επιβάλει την χρήση μεθόδων με χαμηλές απαιτήσεις σε υπολογιστικούς πόρους. Δυστυχώς στην διεθνή αρθρογραφία δεν εμφανίζονται αλγόριθμοι οι οποίοι να συνδυάζουν χαμηλές απαιτήσεις μνήμης και χαμηλή υπολογιστική πολυπλοκότητα. Με αφορμή την παραπάνω παρατήρηση, στα πλαίσια της συγκεκριμένης εργασίας παρουσιάζουμε την FEDRA, έναν αλγόριθμο μείωσης διάστασης με ελάχιστες απαιτήσεις μνήμης και υπολογιστικής ισχύος. Η FEDRA προβάλει τα δεδομένα σε έναν χώρο μικρότερης διάστασης χρησιμοποιώντας ένα σύνολο σημείων αναφοράς. Παράλληλα έχει σημαντικά λιγότερες απαιτήσεις μνήμης συγκριτικά με ανταγωνιστικές μεθόδους ενώ ταυτόχρονα είναι εξίσου γρήγορη χωρίς τα χαρακτηριστικά αυτά να αποβαίνουν εις βάρος της ποιότητας. Στην θεωρητική ανάλυση της προσέγγισής μας προσδιορίζουμε το σφάλμα που υπεισέρχεται λόγω της προβολής και προσδιορίζουμε το άνω και κάτω φράγμα αυτού. Επιπλέον προτείνουμε κάποιες επεκτάσεις του βασικό αλγορίθμου που είναι ιδανικές για υπολογιστές που μπορούν να υποστηρίξουν υπολογιστικά κοστοβόρους αλγόριθμους. Τέλος επικυρώνουμε την ορθότητα των ισχυρισμών μας μέσω εκτενούς πειραματικής αποτίμησης όπου η FEDRA συγκρίνεται με ένα μεγάλο αριθμό γνωστών αλγορίθμων μείωσης διάστασης όπως ο FastMap, ο LMDS, το PCA, το SVD, και τα Random Projections.

Ομοιοκατάταξη++: Αναδιατύπωση ερωτημάτων αναλύωντας συνδέσμους του γράφου των κλικ

Ιωάννης Αντωνέλλης, Hector GarciaMolina, ChiChao Chang

Μελετάμε το πρόβλημα της αναδιατύπωσης ερωτημάτων για πληρωμένες διαφημίσεις. Για την αναδιατύπωση ερωτημάτων βασιζόμαστε σε έναν γράφο που έχει καταγεγραμμένες τις διαφημίσεις που έχουν επιλέξει χρήστες για ερωτήματα που διατύπωσαν στο παρελθόν.
Για κάθε ερώτημα, αρχικά μελετάμε την τεχνική της Ομοιοκατάταξης σαν έναν τρόπο να ανακαλύψουμε άλλα συναφεί με αυτό ερωτήματα: ερωτήματα οι διαφημίσεις των οποίων μπορεί να είναι ενδιαφέρουσες για κάποιον χρήστη.
Υποστηρίζουμε ότι η τεχνική της Ομοιοκατάταξης αποτυγχάνει να αναγνωρίσει την ομοιότητα ερωτημάτων στο συγκεκριμένο πρόβλημα που μελετάμε και παρουσιάζουμε δύο εμπλουτισμένες εκδοχές της τεχνικής της Ομοιοκατάταξης: η πρώτη εκμεταλεύεται υπαρχουσα βάρη στις ακμές του γράφου των κλικ και η δεύτερη εκμεταλεύεται άλλου είδους "μαρτυρίες" που υπάρχουν στον γραφο.
Μελετάμε πειραματικά τις νέες τεχνικές και τις συγκρίνουμε με την τεχνική της Ομοιοκατάταξης, χρησιμοποιώντας πραγματικούς γράφους με κλικ και συλλογές ερωτημάτων απο την μηχανή αναζήτησης της Γιάχου! με βάση δάφορες μετρικές.
Τα αποτελέσματα των πειραμάτων μας δείχνουν ότι οι εμπλουτισμένες τεχνικές μπορούν να αποφέρουν περισσότερες και καλύτερες αναδιατυπώσεις ερωτημάτων.

Αυτοπροσαρμοζόμενη Χρονοδρόμιση Συνδιαλλαγών Παγκόσμιου Ιστού Πληροφοριών

Shenoda Guirguis, Mohamed Sharaf, Πάνος Χρυσάνθης, Αλέξανδρος Λαμπρινίδης, Kirk Pruhs

Η επιτυχία των αλληλεπιδραστικών και δυναμικών συστημάτων βάσεων δεδομένων στο παγκόσμιο ιστό καθορίζεται από το επίπεδο ικανοποίησης των χρηστών για τις υπηρεσίες που λαμβάνουν. Σε αυτά τα

συστήματα, οι ιστοσελίδες που οι χρήστες αναζητούν δημιουργούνται δυναμικά με την εκτέλεση ενός αριθμού συνδιαλλαγών ερωτημάτων σε βάσεις δεδομένων ή συνδιαλλαγών παγκόσμιου ιστού. Σε αυτή την εργασία, μοντελοποιούμε τις συσχετιζόμενες συνδιαλλαγές που δημιουργούν μία ιστοσελίδα ως μια ροή εργασίας (workflow) και ποσοτικοποιούμε την ικανοποίηση των χρηστών με το να συνδέουμε τις συνδιαλλαγές με λανθάνουσες καταληκτικές προθεσμίες (soft-deadlines). Επιπλέον μοντελοποιούμε την σημαντικότητα των συνδιαλλαγών στη δημιουργία μίας ιστοσελίδας με το να συνδέουμε κάθε συνδιαλλαγή με διαφορετικό βάρος. Χρησιμοποιώντας αυτό το πλαίσιο μοντελοποίησης, η επιτυχία ενός συστήματος μπορεί να μετρηθεί με βάση την ελαχιστοποίηση της απόκλισης από την καταληκτική προθεσμία, δηλαδή την καθυστέρηση ή υστέρηση (tardiness), καθώς επίσης την ελαχιστοποίηση της κανονικοποιημένης απόκλισης, δηλαδή, κανονικοποιημένης υστέρησης βάρους (weighted tardiness). Για την αποδοτική δημιουργία δυναμικών ιστοσελίδων, προτείνουμε τον αλγόριθμο ASETS*, που είναι ένας αυτοπροσαρμοζόμενος αλγόριθμος χρονοδρόμησης χωρίς εξωτερικές παραμέτρους. Ο ASETS* αυτόματα προσαρμόζεται όχι μόνο στο φορτίο εργασίας του συστήματος αλλά και στα χαρακτηριστικά των συνδιαλλαγών, όπως οι αλληλοεξαρτήσεις, οι καταληκτικές προθεσμίες και τα βάρη.
Ο ASETS* καθορίζει την προτεραιότητα εκτέλεσης κάθε συνδιαλλαγής έτσι ώστε να ελαχιστοποιήσει την κανονικοποιημένη υστέρηση βάρους. Έχει επίσης την δυνατότητα να συμβιβάζει την βελτιστοποίηση μεταξύ της μέσης και της χειρότερης ανταπόκρισης, αν είναι αυτό επιθυμητό. Η υψηλή απόδοση του ASETS* έχει αναδειχθεί πειραματικά.

Αυτόματη συμπλήρωση για Mashups.

Νεοκλής Πολυζώτης, Tova Milo, Ohad Greenshpan

Ένα mashup είναι μια εφαρμογή του Ιστού που ενοποιεί δεδομένα, λογική υπολογισμού, και στοιχεία διεπαφής από διάφορα συνιστώντα μέρη. Η έννοια του mashup προήλθε από την παρατήρηση ότι όλο και περισσότερες εφαρμογές είναι διαθέσιμες στον Ιστό, οι οποίες πρέπει να συνδυάζονται ώστε να εξυπηρετηθούν οι ανάγκες των χρηστών. Η εργασία μας εισάγει το σύστημα Matchup που υποστηρίζει την γρήγορη ανάπτυξη mashups χρησιμοποιώντας έναν καινοτόμο μηχανισμό αυτόματης συμπλήρωσης. Η κύρια ιδέα είναι ότι τα mashups που αναπτύσσονται από διαφορετικούς χρήστες συνήθως έχουνε παρόμοια χαρακτηριστικά, π.χ., χρησιμοποιούνε παρόμοια μέρη που συνδυάζονται με παρόμοια λογική. Το σύστημά μας εκμεταλλεύεται αυτές τις ομοιότητες για να προτείνει πιθανές συμπληρώσεις (συνθετικά μέρη και τρόπο σύνδεσης) για ένα μερικό mashup. Οι συμπληρώσεις παρουσιάζονται ταξινομημένες στον χρήστη βάσει μιας συνάρτησης χρησιμότητας, η οποία συνυπολογίζει τον βαθμό ταιριάσματος με το μερικό mashup αλλά και την συχνότητα χρήσης των συγκεκριμένων μερών από άλλους χρήστες.

Αναπαριστώντας την Προέλευση RDF Τριάδων με Χρώματα

Γιώργος Φλουρής, Ειρήνη Φουντουλάκη, Παναγιώτης Πεδιαδίτης, Γιάννης Θεοχάρης, Βασίλης Χριστοφίδης

Πρόσφατα η προσπάθεια W3C Linking Open Data έδωσε ώθηση στη δημοσίευση και διασύνδεση μεγάλων όγκων RDF δεδομένων στο Σημασιολογικό Ιστό. Διάφορες οντολογίες και βάσεις γνώσης με εκατομμύρια RDF τριάδες από τη Wikipedia και άλλες πηγές, ως επί το πλείστον προερχόμενων από το e-science, έχουν δημιουργηθεί και δημοσιευθεί. Η καταγραφή πληροφορίας προέλευσης RDF τριάδων συναθροιζόμενων από διαφορετικές ετερογενείς πηγές είναι κρίσιμη ώστε να υποστηριχθούν αποτελεσματικά μηχανισμοί εμπιστοσύνης, ψηφιακά δικαιώματα και πολιτικές απορρήτου. Η διαχείριση της προέλευσης γίνεται ακόμα πιο σημαντική αν θεωρήσουμε όχι μόνο τις άμεσα διατυπωμένες αλλά και τις έμμεσες τριάδες (οι οποίες προκύπτουν με την εφαρμογή των κανόνων εξαγωγής συμπερασμάτων της RDFS) σε συνδυασμό με δηλωτικές γλώσσες επερώτησης και ενημέρωσης RDF

γράφων. Σ' αυτό το άρθρο βασιζόμαστε σε χρωματισμένες RDF τριάδες τις οποίες αναπαριστούμε ως τετράδες για να καταγράψουμε και διαχειριστούμε άμεση πληροφορία προέλευσης.

Ένα παίγνιο διαμόρφωσης συστάδων σε συστήματα ομότιμων βασισμένο στην ποιότητα ανάκλησης ερωτήσεων

Γεωργία Κολωνιάρη, Ευαγγελία Πιτουρά

Σε πολλές εφαρμογές διαμοιρασμού περιεχομένου μεγάλης κλίμακας, οι συμμετέχοντες ομαδοποιούνται με βάση το περιεχόμενο ή τα ενδιαφέροντα τους, δημιουργώντας έτσι συστάδες. Σε αυτήν την εργασία, ασχολούμαστε με την συντήρηση τέτοιων συστάδων υπό την παρουσία ενημερώσεων. Μοντελοποιούμε την εξέλιξη του συστήματος ως ένα παίγνιο, στο οποίο οι κόμβοι καθορίζουν την συμμετοχή τους στις συστάδες βασισμένοι σε μια συνάρτηση ωφέλειας πάνω στην ανάκληση των ερωτήσεων. Οι κόμβοι κατευθύνονται είτε από εγωιστικά είτε από αλτρουιστικά κίνητρα: οι εγωιστικοί κόμβοι στοχεύουν στη βελτίωση της ανάκλησης των δικών τους ερωτήσεων, ενώ οι αλτρουιστικοί κόμβοι στοχεύουν στη βελτίωση της ανάκλησης των ερωτήσεων άλλων κόμβων. Μελετούμε την εξέλιξη τέτοιων συστάδων, τόσο θεωρητικά όσο και πειραματικά κάτω από διάφορες συνθήκες. Δείχνουμε ότι, γενικά, τοπικές αποφάσεις που παίρνονται ανεξάρτητα σε κάθε κόμβο, επιτρέπουν στο σύστημα να προσαρμόζεται σε αλλαγές και να διατηρεί την συνολική ποιότητα ανάκλησης του φορτιού ερωτήσεων.

GRaSP: Γενικευμένη Αναζήτηση Εύρους σε Δίκτυα Ομοτίμων

Μιχαήλ Αργυρίου, Βασίλης Σαμολοαδάς, Σπύρος Μπλάνας

Προτείνουμε ένα πλαίσιο για γενικευμένη αναζήτηση εύρους σε δίκτυα ομοτίμων κόμβων, δομημένων ως tries, όπως είναι το P-Grid. Οι τεχνικές μας βασίζονται σε άγνωστες μέχρι σήμερα ιδιότητες των τυχαίων tries. Αποδεικνύουμε ότι ένα δίκτυο ομοτίμων, παρόμοιο με το P-Grid έχει διάμετρο δρομολόγησης O(log n) με μεγάλη πιθανότητα, καθώς και O(log n) συνωστισμό, ανεξάρτητα από το σχήμα του σχετικού trie. Με βάση αυτές τις ιδιότητες, προτείνουμε το GRaSP, ένα απλό σχήμα που μπορεί να εφαρμοστεί σε κάθε πρόβλημα αναζήτησης εύρους, με κόστος εισαγωγής και ανάκτησης O(log n) βημάτων με μεγάλη πιθανότητα. Εφαρμόζουμε το GRaSP σε δύο προβλήματα αναζήτησης: πολυδιάστατη αναζήτηση σε σημεία και παραλληλόγραμμα, καθώς και σε αναζήτηση τριών πλευρών. Τα πειραματικά μας αποτελέσματα επιβεβαιώνουν ότι το GRaSP προσφέρει εξαιρετική απόδοση αναζήτησης και καλή κλιμακωσιμότητα σε συνθήκες υψηλού φόρτου. Ειδικά στο πρόβλημα αναζήτησης τριών πλευρών, το πλαίσιό μας διακρίνεται ως προς τη βελτίωση της κατανομής φόρτου, εισάγοντας πλεονασμό στην αποθήκευση μέσω κατάλληλης επιλογής του χώρου αναζήτησης.

Υποστήριξη Διαβάθμισης με βάση Προτιμήσεις και Διαφορετικότητα σε Συστήματα Έκδοσης/Συνδρομής

Μαρίνα Δρόσου, Κώστας Στεφανίδης, Ευαγγελία Πιτουρά

Στα συστήματα έκδοσης/συνδρομής, οι χρήστες περιγράφουν τα ενδιαφέροντά τους μέσω συνδρομών και ενημερώνονται όταν νέα, ενδιαφέροντα γεγονότα είναι διαθέσιμα. Συνήθως, σε τέτοια συστήματα, όλες οι συνδρομές θεωρούνται ίσης σημασίας. Ωστόσο, λόγω του μεγάλου όγκου δεδομένων, οι χρήστες μπορεί να λαμβάνουν πολλά γεγονότα. Σε αυτή την εργασία, προτείνουμε ένα μηχανισμό διαβάθμισης γεγονότων με βάση τις προτιμήσεις των χρηστών, έτσι ώστε, μόνο τα πιο ενδιαφέροντα να αποστέλλονται σε αυτούς. Επειδή πολλές φορές αυτά τα γεγονότα μπορεί να είναι πολύ όμοια μεταξύ τους, στοχεύουμε στην αύξηση της διαφορετικότητάς τους. Επιπροσθέτως, εξετάζουμε διαφορετικές πολιτικές για την αποστολή των γεγονότων στους χρήστες. Πιο συγκεκριμένα, μελετάμε (α) μία περιοδική πολιτική, (β) μία πολιτική κυλιόμενου παραθύρου και (γ) μία πολιτική που βασίζεται στην ιστορία των γεγονότων που έχουν ήδη αποσταλεί. Έχουμε υλοποιήσει την προσέγγισή μας

επεκτείνοντας τη SIENA, ένα δημοφιλές σύστημα έκδοσης/συνδρομής, και παραθέτουμε τα αποτελέσματα των πειραμάτων μας.

Αποθήκευση και δεικτοδότηση χωρικών δεδομένων σε συστήματα ομότιμων

Βηρένα Καντερέ, Σπύρος Σκιαδόπουλος, Τίμος Σελλής

Τα συστήματα ομότιμων κόμβων έχουν γίνει πολύ δημοφιλή για την αποθήκευση και ανταλλαγή πληροφοριών με αποκεντρωμένο τρόπο. Αρχικά, η έρευνα επικεντρώθηκε σε συστήματα ομότιμων που φιλοξενούν μονοδιάστατα δεδομένα. Σήμερα, υπάρχει η ανάγκη για εφαρμογές ομότιμων κόμβων που διαχειρίζονται πολυδιάστατα δεδομένα. Η πλειονότητα των προτεινόμενων τεχνικών για πολυδιάστατα δεδομένα βασίζεται είτε στην κατανομή των συγκεντρωτικών ευρετηρίων είτε στη μείωση των διαστάσεων των δεδομένων. Στόχος μας είναι να δημιουργήσουμε εκ του μηδενός μια τεχνική που είναι εγγενώς κατανεμημένη και υποστηρίζει τις πολλαπλές διαστάσεις των δεδομένων χωρίς μείωση. Επικεντρωνόμαστε σε δομημένα συστήματα ομότιμων που μοιράζονται χωρική πληροφορία. Παρουσιάζουμε το σύστημα SpatialP2P, ένα πλαίσιο για εγγενώς αποκεντρωμένη δεικτοδότηση και αναζήτηση χωρικών δεδομένων. Το SpatialP2P υποστηρίζει εφαρμογές ομότιμων συστημάτων όπου χωρική πληροφορία οποιουδήποτε μεγέθους μπορεί να εισαχθεί ή να διαγραφεί και οι ομότιμοι να ενταχθούν ή να αποχωρήσουν δυναμικά. Η προτεινόμενη τεχνική διατηρεί την τοπικότητα και την κατευθυνσιμότητα του χώρου.

Πιστοποίηση της γνησιότητας των αποτελεσμάτων ερωτήσεων συσχέτισης σε βάσεις δεδομένων που έχουν ανατεθεί σε τρίτους

Δημήτρης Παπαδιάς Yin Yang, Σταύρος Παπαδόπουλος, Πάνος Καλνής

Στην περίπτωση που μια βάση δεδομένων έχει ανατεθεί σε τρίτους, ο διακομιστής ερωτήσεων πρέπει να κατασκευάσει μια απόδειξη της ορθότητας των αποτελεσμάτων, που να μπορεί να ελεγχθεί από τον πελάτη χρησιμοποιώντας την υπογραφή του κατόχου των δεδομένων. Οι υπάρχουσες τεχνικές πιστοποίησης της γνησιότητας ασχολούνται με απλές ερωτήσεις διαστήματος σε μια μόνο σχέση χρησιμοποιώντας μια Πιστοποιημένη Δομή Δεδομένων (ΠΔΔ). Στην περίπτωση των ερωτήσεων συσχέτισης, η πιστοποίηση της γνησιότητας των αποτελεσμάτων είναι εγγενώς πιο πολύπλοκη, δεδομένου ότι μόνο οι αρχικές σχέσεις, αλλά όχι ο συνδυασμός τους, υπογράφονται από τον κάτοχο. Σε αυτή την εργασία παραθέτουμε τρεις καινοτόμους αλγορίθμους που διαφοροποιούνται ανάλογα με τις διαθέσιμες ΠΔΔ. (α) Αλγόριθμος AISM: Βασίζεται στη επεξεργασία ερωτήσεων συσχέτισης με ταξινόμηση με συγχώνευση μέσω ευρετηρίου. Προϋποθέτει μόνο μια ΠΔΔ στο πεδίο συσχέτισης. (β) Αλγόριθμος AIM: Προϋποθέτει ότι για κάθε σχέση που μετέχει στην ερώτηση, υπάρχει μια ΠΔΔ στο πεδίο συσχέτισης. Η επεξεργασία της ερώτησης συσχέτισης βασίζεται στην ταξινόμηση με συγχώνευση μέσω ευρετηρίου. (γ) Αλγόριθμος ASM: Δεν απαιτεί καμιά ΠΔΔ. Η επεξεργασία της ερώτησης συσχέτισης γίνεται μέσω ταξινόμησης με συγχώνευση. Τα αποτελέσματα των πειραμάτων δείχνουν ότι οι προτεινόμενες μέθοδοι είναι καλύτερες σε σχέση με τις υπάρχουσες κατά αρκετές τάξεις μεγέθους, σε όλες τις μετρήσεις απόδοσης. Ειδικότερα, οι προτεινόμενες μέθοδοι μεταφέρουν αποτελεσματικά τον φόρτο εργασίας από τον πελάτη στον εξωτερικό διακομιστή στον οποίο έχει ανατεθεί η βάση δεδομένων. Τέλος, επεκτείνουμε τις τεχνικές μας σε σύνθετες ερωτήσεις που συνδυάζουν συσχέτιση πολλών σχέσεων, επιλογές και προβολές.

Υπολογισμός Επερωτήσεων Κορυφογραμμών σε μια Παράλληλη Αρχιτεκτονική

Ακριβή Βλάχου, Χρήστος Δουλκερίδης, Γιάννης Κωτίδης

Πρόσφατα, οι επερωτήσεις κορυφογραμμών (skyline queries) έχουν προσελκύσει το ενδιαφέρον της ερευνητικής κοινότητας διαχείρισης δεδομένων. Τεχνικές διαμέρισης του χώρου αναπαράστασης των δεδομένων, όπως αναδρομική διαμέριση του χώρου, χρησιμοποιούνται για την επεξεργασία επερωτήσεων κορυφογραμμών σε κεντρικοποιημένα, παράλληλα και κατανεμημένα συστήματα. Δυστυχώς, τέτοιου είδους διαμέριση του χώρου με τη χρήση πλέγματος είναι ακατάληλλη στην περίπτωση παράλληλης επεξεργασίας, όπου όλες οι διαμερίσεις επεξεργάζονται την ίδια στιγμή, αφού πολλές διαμερίσεις δε συνεισφέρουν στο τελικό αποτέλεσμα προκαλώντας περιττή επεξεργασία.
Σε αυτό το άρθρο προτείνουμε μια καινοτομική τεχνική διαμέρισης του χώρου βασισμένη σε γωνίες, χρησιμοποιώντας τις υπερσφαιρικές συντεταγμένες των σημείων του συνόλου δεδομένων. Αποδεικνύουμε με τυπικές μεθόδους καθώς και με εξαντλητικά πειράματα ότι η νέα τεχνική είναι κατάλληλη για επεξεργασία κορυφογραμμών σε παράλληλες αρχιτεκτονικές. Διαισθητικά, η τεχνική μας ισοκατανέμει τα σημεία του συνόλου κορυφογραμμής στις διαμερίσεις. Επιπλέον, δείχνουμε ότι διαμερίζοντας τα δεδομένα με βάση τις υπερσφαιρικές συντεταγμένες, κατορθώνουμε να αυξήσουμε το πλήθος των σημείων εντός μιας τυχαίας διαμέρισης που δε χρειάζεται να επεξεργαστούν. Η νέα τεχνική διαμέρισης υπερσκελίζει τα περισσότερα προβλήματα που σχετίζονται με τη διαμέριση με χρήση πλέγματος, μειώνοντας το χρόνο απόκρισης και μοιράζοντας τον υπολογιστικό φόρτο δίκαια. Όπως επιδεικνύεται στην πειραματική μας μελέτη, η τεχνική μας αποδίδει πάντα καλύτερα από τη διαμέριση με χρήση πλέγματος, και προβάλλει ως μια αποδοτική και επεκτάσιμη λύση για τον υπολογισμό επερωτήσεων κορυφογραμμών σε παράλληλες αρχιτεκτονικές.

Αποτίμηση ερωτημάτων προσέγγισης σε συλλογές μονοπατιών

Παναγιώτης Μπούρος, Σπύρος Σκιαδόπουλος, Θοδωρής Δαλαμάγκας, Δημήτρης Σαχαρίδης, Τίμος Σελλής

Πλήθος εφαρμογών σε τομείς όπως η βιοχημεία και τα γεωγραφικά πληροφοριακά συστήματα (GIS), καλούνται να διαχειριστούν και να απαντήσουν ερωτήματα σε μεγάλες συλλογές από ακολουθιακά δεδομένα, αποθηκευμένα ως συλλογές μονοπατιών. Υπάρχει μία πλειάδα ερωτημάτων που μπορούν αποτιμηθούν σε τέτοιου είδους δεδομένα. Η παρούσα δουλειά επικεντρώνεται στα ερωτήματα προσέγγισης: δεδομένης μίας συλλογής μονοπατιών και δύο κόμβων s, t, θέλουμε να διαπιστώσουμε αν υπάρχει μονοπάτι από το s στο t και να αναγνωρίσουμε το μονοπάτι αυτό. Για την αποτίμηση των ερωτημάτων αυτών προτείνουμε τον αλγόριθμο path-first search που αντιμετωπίζει τα μονοπάτια ως πολίτες πρώτης τάξης. Για την περαιτέρω βελτίωση των επιδόσεων των μεθόδων μας, εισαγάγουμε δύο ευρετήρια που αναπαριστούν την πληροφορία πρόσβασης μεταξύ των κόμβων στα μονοπάτια. Επιπλέον, παρουσιάζουμε μεθόδους για την ενημέρωση μιας συλλογής μονοπατιών και των ευρετηρίων της. Τέλος, διεξάγουμε μία εκτενή πειραματική μελέτη που επιβεβαιώνει τα πλεονεκτήματα των μεθόδων μας.

Αποδοτική Αποτίμηση Γενικευμένων Ερωτήσεων Μονοπατιών σε δεδομένα XML

Xiaoying Wu, Στέφανος Σουλδάτος, Δημήτρης Θεοδωράτος, Θοδωρής Δαλαμάγκας, Τίμος Σελλής

Η εύρεση ταιριασμάτων δομικών προτύπων σε δεδομένα XML είναι μια βασική λειτουργία που χρειάζεται στην επεξεργασία ερωτήσεων XML. Οι αλγόριθμοι που έχουν μέχρι στιγμής προταθεί εστιάζουν σε πρότυπα μονοπατιών ή πρότυπα δέντρων. Όμως πρόσφατα έχουν αναδειχτεί γλώσσες ερωτήσεων στις οποίες υπάρχει ευελιξία στον προσδιορισμό της δομής στα πρότυπα μονοπατιών, ώστε να δίνεται η δυνατότητα στο χρήστη πιο ευέλικτης διατύπωσης ερωτήσεων. Στη δημοσίευση αυτή προτείνουμε μεθόδους αποδοτική αποτίμηση γενικευμένων ερωτήσεων προτύπων μονοπατιών σε

δεδομένα XML, που δίνουν τη δυνατότητα μερικού προσδιορισμού δομής. Οι μέθοδοι λαμβάνουν υπόψη και την περίπτωση που υπάρχουν πολλαπλές εμφανίσεις κόμβων με την ίδια ετικέτα σε μια ερώτηση μονοπατιού. Δείχνουμε πώς οι ερωτήσεις προτύπων μονοπατιού με μερική δομή μπορούν να αναπαρασταθούν ως κατευθυνόμενοι ακυκλικοί γράφοι για τους οποίους υπάρχει τοπολογική ταξινόμηση των κόμβων τους. Παρουσιάζουμε τρεις αλγορίθμους αποδοτικής αποτίμησης των ερωτήσεων αυτών. Ο πρώτος αλγόριθμος χρησιμοποιεί μια δομική περίληψη των δεδομένων για την εξαγωγή συνόλου απλών ερωτήσεων προτύπων μονοπατιών που είναι ισοδύναμες με την αντίστοιχη ερώτηση προτύπων μονοπατιού με μερική δομή. Στη συνέχεια αποτιμά τις ερωτήσεις αυτές, επεκτείνοντας υπάρχουσες μεθόδους. Ο δεύτερος αλγόριθμος εξάγει το ελάχιστο ζευγνύον δέντρο από το γράφο της ερώτησης, αποτιμά τα μονοπάτια του δέντρου με χρήση μεθόδων στοίβας και ενώνει τα αποτελέσματα. Ο τρίτος αλγόριθμος χρησιμοποιεί μια τοπολογική ταξινόμηση των κόμβων του γράφου και αποτιμά το ερώτημα ολιστικά με χρήση μεθόδων στοίβας. Τέλος, παρουσιάζουμε αναλυτική πειραματική αποτίμηση των μεθόδων μας και δείχνουμε την ανωτερότητα του τρίτου αλγορίθμου.

EverLast: Μια κατανεμημένη αρχιτεκτονική για τη διατήρηση του Παγκόσμιου Ιστού

Avishek Anand, Srikanta Bedathur, Klaus Berberich, Ralf Schenkel, Χρήστος Τρυφωνόπουλος

Ο Παγκόσμιος Ιστός έχει γίνει μια από τις βασικές πηγές γνώσης που χαρακτηρίζει κάθε έκφανση της σύγχρονης ζωής. Δυστυχώς πολλά από τα δεδομένα του Παγκόσμιου Ιστού είναι εφήμερα, και σύμφωνα με υπολογισμούς περισσότερο από το 50-80% του περιεχομένου αλλάζει μέσα σε σύντομο χρονικό διάστημα. Συνεχίζοντας τις πρωτοπόρες προσπάθειες πολλών εθνικών (ψηφιακών) βιβλιοθηκών, οργανισμοί όπως το International Internet Preservation Consortium (IIPC), το Internet Archive (IA) και το European Archive (EA) εργάζονται για την διατήρηση του συνεχώς μεταβαλλόμενου Παγκόσμιου Ιστού.

Παρόλο που αυτές οι προσπάθειες έχουν εστιάσει σε ζητήματα όπως η μακρόχρονη διατήρηση των δεδομένων του Παγκόσμιου Ιστού, δεν έχουν δώσει την απαιτούμενη προσοχή στην ανάπτυξη μιας υποδομής μεγάλης κλίμακας για την συλλογή, αρχειοθέτηση και ανάλυση των συλλεγόμενων δεδομένων. Βασιζόμενοι σε παρατηρήσεις από την πρόσφατη εργασία μας σε μεθόδους ανάλυσης αρχειοθετημένων κειμένων του Παγκόσμιου Ιστού, προτείνουμε το EverLast, μια κατανεμημένη κλιμακούμενη αρχιτεκτονική για την επόμενη γενιά αρχειοθέτησης κειμένων του Παγκόσμιου Ιστού. Το σύστημά μας βασίζεται σε μια κατανεμημένη αρχιτεκτονική η οποία μπορεί να αναπτυχθεί πάνω από δίκτυα ομότιμων κόμβων επιτρέποντας την ενσωμάτωση των υπαρχόντων προσπαθειών αρχειοθέτησης που έχουν γίνει κυρίως σε εθνικό επίπεδο. Κυρίαρχα χαρακτηριστικά του EverLast είναι η υποστήριξη αναζήτησης και ανάλυσης κειμένων με βάση το χρόνο, και η αρχειοθέτηση κειμένων κατά την πλοήγηση των χρηστών στο Διαδίκτυο. Στην εργασία αυτή περιγράφουμε την συνολική αρχιτεκτονική του συστήματος, και παρουσιάζουμε κάποια αρχικά αλλά υποσχόμενα αποτελέσματα.

Ένα οικονομικό μοντέλο για σύννεφα υπολογιστών διαχείρισης τοπικών αντιγράφων δεδομένων

Βηρένα Καντερέ, Debabrata Dash, Αναστασία Αϊλαμάκη

Το "σύννεφο υπολογιστών", η νέα τάση για τις υποδομές που προσφέρουν διαδικτυακές υπηρεσίες, πρέπει να εγγυάται την ικανοποίηση πολλών χρηστών με ελάχιστες κεφαλαιακές δαπάνες. Σε ένα σύννεφο που προσφέρει υπηρεσίες πάνω σε μεγάλο όγκο δεδομένων που συλλέγονται και ρωτώνται μαζικά, όπως επιστημονικά δεδομένα, οι χρήστες πληρώνουν για τις υπηρεσίες που δέχονται. Το σύννεφο υποστηρίζει την προσωρινή αποθήκευση των δεδομένων, ώστε να μπορεί να παρέχει

υπηρεσίες ερωτήσεων μεγάλης ποιότητας. Οι χρεώσεις των χρηστών που ζητούν να εκτελεστούν τα ερωτήματά τους καλύπτουν το κόστος και τη συντήρηση των υποδομών του σύννεφου. Η πρόκληση έγκειται στην παροχή αποδοτικών και οικονομικών υπηρεσιών διατηρώντας παράλληλα το σύννεφο κερδοφόρο. Στην εργασία αυτή προτείνουμε ένα οικονομικό μοντέλο για ένα αυτο-συντονιζόμενο σύννεφο που παρέχει υπηρεσίες ερωτήσεων πάνω σε επιστημονικά δεδομένα. Η πολιτική της προτεινόμενης οικονομίας ενθαρρύνει την υψηλή ποιότητα υπηρεσιών ερωτήσεων για όλους τους χρήστες, αλλά συγχρόνως προστατεύει το κέρδος του σύννεφου. Ακόμη, προτείνουμε ένα μοντέλο κόστους των υπηρεσιών που λαμβάνει υπόψη όλες τις πιθανές δαπάνες για την εκτέλεση των ερωτημάτων αλλά και δαπάνες υποδομής του σύννεφου. Η πειραματική μελέτη αποδεικνύει ότι η προτεινόμενη λύση είναι βιώσιμη για ποικιλία ερωτημάτων και δεδομένων.

Top-k Επικρατέστερες Υπηρεσίες Ιστού με τη Χρήση Πολλαπλών Κριτηρίων Ταιριάσματος

Δημήτριος Σκούτας, Δημήτρης Σαχαρίδης, Άλκης Σιμιτσής, Βηρένα Καντερέ, Τίμος Σελλής

Καθώς μεταβαίνουμε από έναν Ιστό δεδομένων σε έναν Ιστό Υπηρεσιών, η προσθήκη στις σημερινές μηχανές αναζήτησης του Ιστού δυνατοτήτων για την αποτελεσματική και αποδοτική επιλογή υπηρεσιών καθίσταται σημαντικό πρόβλημα. Τυπικά, ο βαθμός ταιριάσματος ανάμεσα σε μία προσφερόμενη και μία ζητούμενη υπηρεσία καθορίζεται υπολογίζοντας ένα συνολικό σκορ, το οποίο συναθροίζει τους επιμέρους βαθμούς ταιριάσματος μεταξύ των διαφόρων παραμέτρων αυτών των υπηρεσιών. Αυτές οι μέθοδοι χαρακτηρίζονται από δύο βασικά μειονεκτήματα. Πρώτον, δεν υπάρχει ένα συγκεκριμένο κριτήριο ταιριάσματος που να είναι βέλτιστο για τον υπολογισμό της ομοιότητας μεταξύ των συγκρινόμενων παραμέτρων. Αντιθέτως, υπάρχουν πολλές προσεγγίσεις, που ποικίλουν από τη χρήση κριτηρίων ομοιότητας από το χώρο της Ανάκτησης Πληροφορίας μέχρι τη χρήση σημασιολογικού ταιριάσματος με κανόνες λογικής. Δεύτερον, η συνάθροιση των επιμέρους βαθμών ομοιότητας οδηγεί σε σημαντική απώλεια πληροφορίας. Καθώς δεν υπάρχει ένας ενιαίος τρόπος για τη στάθμιση αυτών των επιμέρους βαθμών, οι υπάρχουσες μέθοδοι ακολουθούν συνήθως ένα απαισιόδοξο σενάριο. Κατά συνέπεια, πολλές υπηρεσίες, για παράδειγμα εκείνες που έχουν μία μόνο παράμετρο χωρίς ταίριασμα, μπορεί να αποκλειστούν από τα αποτελέσματα αν και ενδέχεται να αποτελούν καλές εναλλακτικές επιλογές. Στην εργασία αυτή, παρουσιάζουμε μία μέθοδο που αντιμετωπίζει και τα δύο αυτά μειονεκτήματα. Δεδομένης μίας ζητούμενης υπηρεσίας, εισάγουμε ένα αντικειμενικό κριτήριο το οποίο αναθέτει έναν βαθμό επικράτησης σε κάθε διαθέσιμη υπηρεσία. Ο βαθμός αυτός λαμβάνει υπόψη όλα τα διαθέσιμα κριτήρια για κάθε παράμετρο της ζητούμενης υπηρεσίας. Μελετάμε τρεις διαφορετικούς ορισμούς του βαθμού επικράτησης και παρουσιάζουμε αποδοτικούς αλγορίθμους που επιλέγουν τις k επικρατέστερες υπηρεσίες σε κάθε περίπτωση. Εκτενής πειραματική μελέτη τόσο σε πραγματικά όσο και συνθετικά δεδομένα επιδεικνύει την αποτελεσματικότητα και την αποδοτικότητα της προτεινόμενης τεχνικής και των αλγορίθμων.

Minersoft: Μηχανή Αναζήτησης για Πόρους Λογισμικού σε μεγάλης κλίμακας Υποδομές Πλέγματος

Μάριος Δ. Δικαιάκος, Αστέριος Κατσιφοδήμος, Γιώργος Πάλλης

Σε αυτή την εργασία ερευνούμε το πρόβλημα της υποστήριξης αναζήτησης με λέξεις κλειδιά για τον εντοπισμό πόρων λογισμικού οι οποίοι είναι εγκατεστημένοι σε μία μεγάλης κλίμακας υποδομής πλέγματος. Παρουσιάζουμε έναν ιχνηλάτη πλέγματος, γνωστός ως Minersoft, ο οποίος επισκέπτεται διάφορες τοποθεσίες πλέγματος, εντοπίζει και κατηγοριοποιεί τα αρχεία λογισμικού και ανακαλύπτει τις μεταξύ τους συσχετίσεις. Τα αποτελέσματα της ιχνηλασίας του Minersoftαναπαρίστανται από ένα

γράφο. Ένα σύνολο από αλγορίθμους ανάκτησης πληροφορίας χρησιμοποιούνται ώστε να εμπλουτιστεί ο γράφος με συσχετίσεις που βασίζονται στη δομή και στο περιεχόμενο των αρχείων λογισμικού. Παρουσιάζουμε μία μελέτη αξιολόγησης της προσέγγισης μας χρησιμοποιώντας δεδομένα που εξάγονται από μία πραγματική υποδομή πλέγματος, το EGEE. Τα πειραματικά αποτελέσματα έδειξαν πως το Minersoft αποτελεί ένα ισχυρό εργαλείο για την αναζήτηση πόρων λογισμικού σε μεγάλης κλίμακας υποδομές πλέγματος επιτυγχάνοντας υψηλές αποδόσεις.

Συχνά εμφανιζόμενα στοιχεία σε δεδομένα συνεχούς ροής: Πειραματική αξιολόγηση των αλγορίθμων προηγμένης τεχνολογίας

Nishad Manerikar, Θέμης Παλπάνας

Το πρόβλημα της αναγνώρισης των συχνά εμφανιζόμενων στοιχείων σε δεδομένα συνεχούς ροής είναι κοινό σε πολλές εφαρμογές και σε διαφορετικούς τομείς. Αρκετοί αλγόριθμοι, βασιζόμενοι σε διάφορες τεχνολογίες, έχουν προταθεί για την επίλυση αυτού του προβλήματος. Σε αυτή την εργασία κάνουμε μια συνοπτική ανασκόπηση αυτών των αλγορίθμων, και παρουσιάζουμε τα αποτελέσματα μιας ενδελεχούς συγκριτικής πειραματικής μελέτης για την αποδοτικότητά τους. Εξετάσαμε τους αλγόριθμους με συνθετικά και πραγματικά δεδομένα, χρησιμοποιώντας μία κοινή πειραματική πλατφόρμα, και μελετήσαμε την απόδοσή τους αναφορικά με διάφορες παραμέτρους (σχετικές με τον τρόπο λειτουργίας των αλγορίθμων, αλλά και με τα χαρακτηριστικά των πειραματικών δεδομένων). Παρουσιάζουμε τα αποτελέσματα των πειραμάτων και την εμπειρία που αποκομίσαμε απο αυτά.

Προτάσεις ερωτημάτων για διαδραστική εξερεύνηση βάσεων δεδομένων

Γκλόρια Χατζοπούλου, Μαγδαληνή Ειρηνάκη, Νεοκλής Πολυζώτης

Τα συστήματα σχεσιακών βάσεων δεδομένων γίνονται συνεχώς πιο δημοφιλή στην επιστημονική κοινότητα καθώς υποστηρίζουν διαδραστική εξερεύνηση μεγάλου όγκου δεδομένων. Σε ένα τέτοιο σενάριο, οι χρήστες χρησιμοποιούν μια διεπαφή (συνήθως σε ιστοσελίδα) για να στείλουν μια σειρά από SQL ερωτήματα που στοχεύουν στην ανάλυση δεδομένων και στην εξόρυξη χρήσιμης πληροφορίας. Οι χρήστες μπορεί να μην έχουν τις απαραίτητες γνώσεις για το πως να ξεκινήσουν την εξερεύνηση. Άλλες φορές, οι χρήστες μπορεί να παραβλέψουν ερωτήματα που θα ανακτούσαν σημαντική πληροφορία. Θέλοντας να βοηθήσουμε τους χρήστες σε αυτό το πλαίσιο, αντλούμε έμπνευση από τα συστήματα εξατομίκευσης στο διαδίκτυο ώστε να προτείνουμε εξατομικευμένα ερωτήματα στο χρήστη. Η ιδέα μας στηρίζεται στο να παρακολουθούμε τα ερωτήματα του χρήστη, να αναγνωρίζουμε τα τμήματα της βάσης που μπορεί να είναι ενδιαφέροντα για τη συγκεκριμένη ανάλυση δεδομένων, και να προτείνουμε ερωτήματα που θα ανακτούν σχετικά δεδομένα. Παρουσιάζουμε μια αρχική πειραματική μελέτη βασισμένη σε ερωτήματα πραγματικών χρηστών που δείχνει ότι το σύστημά μας μπορεί να προτείνει χρήσιμα ερωτήματα.

Εξαρτήσεις Δεικτών στην Φυσική Σχεδίαση Σχεσιακών Βάσεων Δεδομένων

Νεοκλής Πολυζώτης, Karl Schnaitter, Lise Getoor

Ένα από τα κύρια καθήκοντα ενός διαχειριστή μιας βάσης είναι η επιλογή των κατάλληλων δεικτών σε σχέση με το τρέχον φόρτο ερωτημάτων. Προσπαθώντας να απλοποιήσουν αυτήν την εργασία, τα μοντέρνα συστήματα μπορούν να προτείνουν δείκτες αναλύοντας ένα αντιπροσωπευτικό σύνολο ερωτημάτων. Ο διαχειριστής παραμένει ο τελικός υπεύθυνος, επιλέγοντας ποιοι δείκτες θα υλοποιηθούν και πότε. Αυτή η απόφαση απαιτεί γνώση για το όφελος των προτεινόμενων δεικτών, κάτι το οποίο είναι δύσκολο να κατανοηθεί όταν υπάρχουνε αλληλεπιδράσεις ή εξαρτήσεις μεταξύ των δεικτών. Δυστυχώς, κανένα εμπορικό σύστημα δεν προσφέρει αυτή τη πληροφορία μαζί με τους προτεινόμενους δείκτες.

Παρακινούμενοι από αυτή την έλλειψη, προτείνουμε μια μεθοδολογία και σχετικά εργαλεία που βοηθούν τον διαχειριστή να κατανοήσει τις εξαρτήσεις των προτεινόμενων δεικτών. Τυποποιούμε την έννοια της εξάρτησης και αναπτύσσουμε έναν αλγόριθμο υπολογισμού των εξαρτήσεων μέσα σε ένα σύνολο δεικτών. Παρουσιάζουμε πειραματικά αποτελέσματα με μία πρωτότυπη υλοποίηση πάνω από το σύστημα IBM DB2 που καταδεικνύουν την αποδοτικότητα της προσέγγισής μας. Επίσης, περιγράφουμε δύο βοηθητικά εργαλεία για διαχειριστές που χρησιμοποιούν την πληροφορία των εξαρτήσεων. Το πρώτο οπτικοποιεί τις εξαρτήσεις βάσει ενός διαμερισμού των δεικτών σε υποσύνολα τα οποία δεν αλληλεπιδρούν, και το δεύτερο υπολογίζει ένα πρόγραμμα κατασκευής των δεικτών κατά πολλαπλές περιόδους συντήρησης της βάσης, ώστε να μεγιστοποιηθεί το συνολικό όφελος για τα ερωτήματα. Και στις δύο περιπτώσεις παρουσιάζουμε ισχυρά θεωρητικά αποτελέσματα που δείχνουν ότι η γνώση των εξαρτήσεων επιτρέπει βελτιωμένη λειτουργικότητα.

Τεχνικές Επεξεργασίας Επερωτήσεων για Οδηγούς Μνήμης Flash

Δημήτρης Τσιρογιάννης, Σταύρος Χαριζόπουλος, Mehul Shah, Janet Wiener, Goetz Graefe

Οι οδηγοί μνήμης Flash (solid state drives - SSD) πραγματοποιούν τυχαίες προσπελάσεις δεδομένων περισσότερο από 100 φορές πιο γρήγορα από τους παραδοσιακούς μαγνητικούς σκληρούς δίσκους, ενώ προσφέρουν συγκρίσιμο εύρος μεταφοράς δεδομένων κατά τη διάρκεια σειριακής ανάγνωσης και γραφής. Παράλληλα, μειώνουν εξαιρετικά την κατανάλωση ενέργειας. Με βάση αυτά τα χαρακτηριστικά, οι οδηγοί μνήμης Flash καταφέρνουν να επιταχύνουν εφαρμογές λογισμικού και ως εκ τούτου, αναμένεται να αντικαταστήσουν τους παραδοσιακούς σκληρούς δίσκους ως κύριο μέσο μόνιμης αποθήκευσης πληροφορίας σε μεγάλα κέντρα δεδομένων.

Παρόλο που είναι σαφές ότι οι οδηγοί μνήμης Flash μπορούν να ωφελήσουν άμεσα εφαρμογές λογισμικού που βασίζονται σε ανάγνωση δεδομένων με τυχαία προσπέλαση, η αποδοτική χρήση τους σε εφαρμογές που στηρίζονται σε σειριακές προσπελάσεις αποτελεί πρόκληση. Οι εφαρμογές βάσεων δεδομένων, και ειδικά αυτές που εκτελούν εκτενείς επερωτήσεις ανάλυσης δεδομένων, στηρίζονται σε μηχανισμούς που έχουν σχεδιαστεί με βάση τη διαφορετική ταχύτητα μεταξύ τυχαίας και σειριακής προσπέλασης, ενώ οι σχετικοί αλγόριθμοι δίνουν έμφαση στη σειριακή προσπέλαση δεδομένων που αποθηκεύονται στα μόνιμα μέσα αποθήκευσης πληροφορίας.

Σε αυτήν την εργασία ερευνούμε δομές δεδομένων και αλγορίθμους που εκμεταλλεύονται τις γρήγορες τυχαίες αναγνώσεις των οδηγών μνήμης Flash για να επιταχύνουν τις διαδικασίες επιλογής, προβολής, και ζεύξης κατά την επεξεργασία επερωτήσεων σε σχεσιακές βάσεις δεδομένων. Αρχικά αποδεικνύουμε πως οποιαδήποτε διάταξη δεδομένων παρέχει φυσικό διαχωρισμό (σε στήλες) των τιμών των διαφόρων πεδίων, ελαττώνει τον όγκο των δεδομένων προς ανάγνωση κατά τη διάρκεια επερωτήσεων επιλογών και προβολών. Κατόπιν, παρουσιάζουμε τον αλγόριθμο FlashJoin, ένα γενικευμένο αλγόριθμο ζεύξης ο οποίος υποστηρίζει διασωληνώσεις και ελαχιστοποιεί τις προσβάσεις σε αρχικά και ενδιάμεσα σχεσιακά δεδομένα. Ο αλγόριθμος FlashJoin ελαττώνει σημαντικά τις απαιτήσεις σε κυρίως μνήμη καθώς και το κόστος προσπέλασης αποθηκευμένων δεδομένων για οποιαδήποτε ζεύξη μέσα σε μία επερώτηση. Υλοποιήσαμε τις προτεινόμενες τεχνικές τροποποιώντας τους μηχανισμούς της Postgres και πειραματιστήκαμε με έναν οδηγό μνήμης Flash υψηλής απόδοσης. Οι τεχνικές μας βελτίωσαν τους χρόνους εκτέλεσης επερωτήσεων μέχρι και 6 φορές για επερωτήσεις που κυμαίνονται από απλές σειριακές προσπελάσεις δεδομένων και απλές ζεύξεις μέχρι πλήρεις επερωτήσεις TPC-H.

Shore-MT: Ένα κλιμακούμενο σύστημα αποθήκευσης δεδομένων για την εποχή των πολυπύρηνων επεξεργαστών

Ιπποκράτης Πανδής, Ryan Johnson, Νίκος Χαρδαβέλλας, Αναστασία Αϊλαμάκη, Babak Falsafi

Ανέκαθεν τα συστήματα αποθήκευσης δεδομένων ήταν ικανά να εξυπηρετούν παράλληλα πολλαπλά αιτήματα. Ωστόσο, μέχρι πρότινος οι υπολογιστές είχαν μόνο ένα μικρό πλήθος μονοπύρηνων επεξεργαστών. Με αποτέλεσμα μόνο ένα μικρό πλήθος νημάτων να έχουν πραγματικά παράλληλη πρόσβαση στις εσωτερικές δομές του συστήματος. Αυτό επέτρεπε στα συστήματα αποθήκευσης να χρησιμοποιούν μη κλιμακούμενες προσεγγίσεις στους εσωτερικούς μηχανισμούς τους χωρίς κυρώσεις στην απόδοση τους. Όμως, με την εμφάνιση των πολυπύρηνων επεξεργαστών η κατάσταση αυτή μεταβάλλεται. Όλο και περισσότερα νήματα μπορούν να τρέχουν παράλληλα, πιέζοντας την ικανότητα κλιμάκωσης του συστήματος αποθήκευσης. Τα συστήματα που έχουν σχεδιαστεί να έχουν υψηλή απόδοση σε έναν περιορισμένο αριθμό πυρήνων δεν έχουν εξασφαλισμένη αντίστοιχη υψηλή απόδοση σε έναν ακόμα μεγαλύτερο αριθμό πυρήνων, εξαιτίας μη αναμενόμενων εμποδίων στην κλιμάκωση.
Σε αυτή την εργασία μετράμε την απόδοση τεσσάρων δημοφιλών συστημάτων αποθήκευσης ανοιχτού λογισμικού (Shore, BerkeleyDB, MySQL, και PostgreSQL) σε ένα σύγχρονο πολυπύρηνο μηχάνημα. Διαπιστώνουμε ότι όλες οι μηχανές αποθήκευσης παρουσιάζουν προβλήματα κλιμάκωσης και συνοπτικά εξετάζουμε τις στενώσεις τους. Στη συνέχεια παρουσιάζουμε το Shore-MT, μια πολυ-νηματική και ιδιαίτερα κλιμακούμενη έκδοση του Shore, το οποίο αναπτύξαμε προσδιορίζοντας και επιτυχώς αφαιρώντας τις εσωτερικές του στενώσεις. Σε σύγκριση με τα άλλα συστήματα αποθήκευσης δεδομένων το Shore-MT είναι πιο κλιμακούμενο και έχει 2 με 4 φορές μεγαλύτερη απόλυτη απόδοση. Δείχνουμε ότι οι σχεδιαστές των συστημάτων θα πρέπει να εστιάζουν στην κλιμάκωση παρά στις επιδόσεις σειριακών υπολογισμών, και παρουσιάζουμε σημαντικές αρχές για το σχεδιασμό κλιμακούμενων μηχανών αποθήκευσης, με πραγματικά παραδείγματα από την ανάπτυξη του Shore-MT.

Έξυπνη ανασύσταση πλειάδων σε Column-stores.

Στράτος Ιδρέος, Martin Kersten, Stefan Manegold

Τα Column-stores αποθηκεύουν και επεξεργάζονται τα δεδομένα ανά στήλη. Έχουν πολλά πλεονεκτήματα, π.χ. μειωμένο I/O. Το μειονέκτημα είναι ότι χρειάζεται δυναμική ανασύσταση πλειάδων. Το ιδανικό σχήμα είναι να διατηρούμε πολλαπλά στιγμιότυπα κάθε σχέσης, όπου το κάθε στιγμιότυπο να είναι ταξινομημένο με βάση μια από τις στήλες. Αυτό όμως προϋποθέτει γνώση του φόρτου εργασίας, διαθέσιμο χρόνο για την προετοιμασία του σχήματος και ένα περιβάλλον χωρίς ενημερώσεις. Σε αυτή την εργασία προτείνουμε ένα νέο σχήμα, partial sideways cracking, το οποίο καταφέρνει τα ίδια πλεονεκτήματα με την πλήρη ταξινόμηση χωρίς τους παραπάνω περιορισμούς. Λειτουργεί για δυναμικά περιβάλλοντα χωρίς ελεύθερο χρόνο και με συχνές ενημερώσεις.

DEMOS

Ένα εργαλείο για την ανακάλυψη αντιστοιχίσεων μεταξύ αποκαλυπτόμενων σχημάτων

Ο αποκεντρωμένος συντονισμός της ανταλλαγής πληροφορίας σε ευρεία κλίμακα πρέπει να αντιμετωπίσει την ετερογένεια των διαμοιραζόμενων δεδομένων. Τα διαμοιραζόμενα δεδομένα έχουν διαφορετικές μορφές, δηλαδή, δομές, τιμές και σχήματα. Η αναζήτηση αυτών των δεδομένων σε διαφορετικές πηγές συνεπάγεται τεχνικές που μπορούν να γεφυρώσουν τους μορφότυπους των δεδομένων. Ορισμένες από αυτές τις τεχνικές ασχολούνται με την παραγωγή αντιστοιχίσεων για τα σχήματα των δεδομένων. Οι υπάρχουσες τεχνικές καλύπτουν συμπληρωματικές πτυχές του

προβλήματος της αντιστοίχισης σχημάτων. Οι πιο σημαντικές από αυτές ασχολούνται με την παραγωγή όλων των δυνατών αντιστοιχίσεων για ένα ζευγάρι σχημάτων, συνοδεύουν τις αντιστοιχίσεις με σημασιολογία και προσαρμόζουν τις σωστές αντιστοιχίσεις καθώς τα σχήματα εξελίσσονται. Προς την κατεύθυνση αυτή αναπτύσσουμε μια τεχνική που ανακαλύπτει αντιστοιχίσεις καθώς τα σχήματα αυτόνομων πηγών αποκαλύπτονται σταδιακά. Σε αυτή την επίδειξη παρουσιάζουμε ένα νέο πρωτότυπο εργαλείο που υλοποιεί αυτή την τεχνική. Η ανακάλυψη αντιστοιχίσεων είναι σχηματο-κεντρική και ενσωματώνει νέα σημασιολογία καθώς αυτή αποκαλύπτεται. Το εργαλείο συλλέγει την εμπειρία ανακάλυψης αντιστοιχίσεων και την επαναχρησιμοποιεί. Το εργαλείο συνεργάζεται με τον χρήστη ο οποίος καθοδηγεί τη διαδικασία αναζήτησης αντιστοιχίσεων. Η επίδειξη παρουσιάζει σενάρια εφαρμογής του εργαλείου που αποδεικνύουν την καταλληλότητά και την αποτελεσματικότητά του σε ρεαλιστικές καταστάσεις της ενσωμάτωσης και ανταλλαγής δεδομένων μεταξύ ετερογενών αυτόνομων πηγών.

KSpot: Αποδοτική Παρακολούθηση των Κορυφαίων-Κ Απαντήσεων σε Ασύρματα Δίκτυα Αισθητήρων

Αυτή η επίδειξη παρουσιάζει το σύστημα KSpot, το οποίο μπορεί να χρησιμοποιηθεί για την αποδοτική παρακολούθηση των Κορυφαίων-Κ (Top-K) απαντήσεων σε μία επερώτηση Q σε ένα ασύρματο δίκτυο αισθητήρων. Το KSpot αποτελείται από ένα σύστημα κατάταξης και μία γραφική διεπιφάνεια χρήσης. Το σύστημα κατάταξης αξιοποιεί κατανεμημένους αλγόριθμους επεξεργασίας Κορυφαίων-Κ απαντήσεων έτσι ώστε να ελαχιστοποιήσει την κατανάλωση των περιορισμένων πόρων που έχει ένα ασύρματο δίκτυο αισθητήρων επιτυγχάνοντας έτσι την μακροζωία του. Επιπρόσθετα, το KSpot είναι φιλικό και προσαρμοστικό προς το χρήστη παρέχοντας ένα σύστημα διαπροσωπείας όπου ο χρήστης ορίζει με δηλωτικό τρόπο επερωτήσεις τύπου SQL παραλαμβάνοντας τα αποτελέσματα γραφικά αντί υπο μορφή πίνακα.

Για την επίδειξη της εφαρμοσιμότητας του συστήματος κατά τη διάρκεια του συνεδρίου, θα παρουσιάζουμε συνεχώς τους χώρους με το ψηλότερο δείκτη θορύβου έτσι ώστε να γνωρίζουν οι παρευρισκόμενοι που υπάρχει περισσότερη κινητικότητα και πιο ζωηρές συζητήσεις. Το σύστημα θα επιτρέπει επίσης στους χρήστες να προσαρμόσουν τις παραμέτρους του συστήματος (π.χ., την παράμετρο Κ, το είδος μέτρησης, κτλ.) Τέλος, θα επιδείξουμε επίσης τις αποταμιεύσεις ενέργειας τις οποίες επιτυγχάνει το KSpot μέσω ενός υποσυστήματος που παρέχεται από τη γραφική διεπιφάνεια του συστήματος μας.